LOW-LIGHT IMAGE ANALYSIS AND CONTRAST ENHANCEMENT USING GAUSSIAN PROCESS

LOH YUEN PENG

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

LOW-LIGHT IMAGE ANALYSIS AND CONTRAST ENHANCEMENT USING GAUSSIAN PROCESS

LOH YUEN PENG

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Loh Yuen Peng

Registration/Matrix No.: WHA130051

Name of Degree: Doctor of Philosophy

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Low-light Image Analysis and Contrast Enhancement using Gaussian Process

Field of Study: Image Processing (Computer Science)

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:
Designation:

LOW-LIGHT IMAGE ANALYSIS AND

CONTRAST ENHANCEMENT USING GAUSSIAN PROCESS

ABSTRACT

Low-light is an inescapable element in daily surroundings that greatly affects the efficiency of human vision. However, current studies in low-light fundamentally lack an in-depth understanding of natural vision in low-light that would strengthen the development of effective algorithms. This has subsequently restricted the development of well-rounded systems that would aid in low-light environments, such as assistive systems, surveillance, and autonomous car driving. Therefore, this thesis aims to study low-light image data to gain a better understanding of their characteristics, and then based on this understanding, investigate a computer vision solution that would pave the way for the advancement of future assistive systems to operate in low-light conditions. An obvious challenge faced in this study is the lack of a go-to database in this domain, hence led to the first contribution that is a collection of 7,363 low-light images gathered from multiple sources, with 12 object classes annotation in order to facilitate the analysis for the purpose of applications. From this dataset, it was found that low-light environments can be categorized into 10 illumination types, each with different global and local characteristics that could have different impact on a system. The second contribution is an in-depth analysis of the collected data, specifically, by studying the global and local pixel intensities, followed by the performance and visualizations of hand-crafted and learned features. It is found that characteristics of the low-light pixel intensities provide a great challenge to algorithms. The design of conventional hand-crafted features are greatly rooted to the behaviors of bright environments, that they are unable to adequately address noise and lack of details accompanying low-light images. Whereas, learned features revealed that the same object yields amply different features in bright and low-light conditions, and irregular illumi-

nation greatly challenges the attention of the said features. These insights prompt the third contribution, to propose a low-light contrast enhancement algorithm that is not only able to improve the visibility but more importantly to reveal informative features to assist high level applications. To this end, the Gaussian Process is studied as the contrast enhancement approach to model the complexity of the local luminance variations, the primary difficulty in low-light images. Experimental results show that the proposed method outperforms the state-of-the-art in the common visual quality measure, the peak signal-to-noise ratio (PSNR) by 1.17dB. Additionally, novel information retrieval measurements are proposed to better evaluate the usefulness of enhancement algorithms in applications, namely the local features matching and l_1-norm distance measure of intensity histogram. Both of which the proposed method outperforms the state-of-the-art method by a large margin, signifying the applicability of the proposal to support computer vision systems. As a whole, the contributions of this study will push forward the advancement of computer vision towards practicality in low-light environments which will be particularly valuable in the development of assistive and surveillance systems that ensure the quality of life and safety of the public.

Keywords: Low-light, image analysis, image enhancement, gaussian process.

ANALISA IMEJ CAHAYA RENDAH DAN

PENINGKATAN KONTRAS MENGGUNAKAN PROSES GAUSSIAN

ABSTRAK

Cahaya rendah merupakan unsur semulajadi yang tidak dapat dielakkan dalam persekitaran dan ia menjejaskan kecekapan penglihatan manusia. Walaupun demikian, penyelidikan kini pada asasnya kurang pemahaman yang mendalam mengenai penglihatan semulajadi dalam keadaan kurang cahaya yang mungkin membantu dalam pembangunan algorithma yang berkesan. Akibatnya, pembangunan sistem seperti sistem bantuan, pengawasan, dan kenderaan autonomi yang cekap dalam keadaan kurang cahaya adalah terhad. Oleh sebab itu, tesis ini bertujuan menyelidik data imej yang ditangkap dalam keadaan cahaya rendah untuk meningkatkan pemahaman terhadap ciri-cirinya. Kemudiannya, pengetahuan ini digunakan untuk membina sistem penglihatan komputer yang membantu kemajuan sistem bantuan yang boleh beroperasi dalam cahaya rendah. Cabaran yang terbesar dalam penyelidikan ini ialah kekurangan sebuah pangkalan data, oleh itu, sumbangan pertama kajian ini adalah pengumpulan imej cahaya rendah sebanyak 7,363 yang diambil dari pelbagai sumber dengan label bagi 12 kelas benda bagi kerja analisa terhadap aplikasi yang seterusnya. Dalam dataset ini didapati bahawa suasana kurang cahaya boleh dibahagi kepada 10 jenis pencahayaan yang mempunyai sifat-sifat sejagat dan setempat yang tersendiri dan menghasilkan pengaruh yang berlainan terhadap sesebuah sistem. Oleh sedemikian, sumbangan kedua ialah analysis yang terperinci mengenai data yang telah dikumpulkan, khususnya, dengan meneliti keamatan piksel secara menyeluruh dan tempatan, disamping prestasi dan vektor sifat yang diperolehi menggunakan algorithma rekaan-tangan dan juga pembelajaran mesin melalui teknik pengambaran. Didapati bahawa ciri-ciri yang terkandung dalam piksel keamatan cahaya rendah adalah amat mencabar untuk ditangani oleh algorithma. Kebanyakan kaedah-kaedah yang

mengunakan rekaan-tangan lazim di reka bentuk mengikut unsur-unsur imej yang ditangkap dalam persekitaran yang mempunyai cahaya mencukupi, oleh itu kaedah-kaedah tersebut tidak dapat mengendalikan hingar dan kekurangan butiran yang biasa mengiringi imej kurang cahaya. Sebaliknya, pembelajaran mesin mendedahkan bahawa vektor sifat yang diperoleh dalam keadaan cahaya mencukupi dan keadaan kurang cahaya adalah berbeza, dan ketidaksekataan cahaya adalah amat mencabar untuk ditangani. Penemuan-penemuan tersebut menuju kajian ini ke sumbangan yang ketiga, iaitu mencadangkan teknik peningkatan kontras imej kurang cahaya yang bukan sahaja memperbaiki keterlihatan kandungan imej, tetapi yang lebih pentingnya boleh mendedahkan sifat-sifat yang berguna bagi aplikasi tahap tinggi. Untuk melakukan sedemikian, teknik Gaussian Process telah dikaji sebagai kaedah peningkatan kontras untuk pemodelan pencahayaan tempatan yang tidak serata, cabaran utama dalam imej kurang cahaya. Keputusan eksperimen menunjukkan kaedah yang dicadangkan mengatasi kaedah-kaedah yang terbaik dan terkini dalam penilaian kualiti lazim, peak signal-to-noise ratio (PSNR) sebanyak 1.17dB. Tambahan lagi, cara penilaian baru yang berdasarkan pemulihan maklumat juga diperkenalkan bagi menilai kegunaan teknik peningkatan kontras terhadap aplikasi, iaitu local features matching dan l1-norm distance measure of intensity histogram. Teknik yang dicadangkan juga mengatasi kaedah-kaedah terbaik sedia ada bagi kedua-dua penilai baru ini dengan margin yang besar, menandakan kebolehgunaan cadangan ini untuk menyokong sistem penglihatan komputer. Keseluruhannya, sumbangan-sumbangan dari kajian ini akan memajukan sistem penglihatan komputer ke arah keberkesanan dalam persekitaran cahaya rendah yang adalah amat bernilai bagi pembangunan sistem bantuan dan pengawasan untuk menjamin mutu kehidupan dan keselamatan orang awam.

Kata kunci: Cahaya rendah, analisa imej, peningkatan imej, gaussian process.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincerest gratitude towards Dr. Chee Seng Chan for being my supportive supervisor throughout the years of my PhD research. For without his guidance and enthusiasm, the completion of my study would not be possible.

Secondly, I would like to extend my appreciation to Prof. Xuefeng Liang for giving me the opportunity to visit and conduct research work in Prof. Takatsune Kumada's laboratory in Kyoto University. The inspiration and experiences with the fellow members of the lab have been indespensible to my progress.

My most heartfelt thanks goes to my fellow lab members past and present. Dr. Mei Kuan Lim, Dr. Sze Ling Tang, Dr. Chern Hong Lim, Dr. Wai Lam Hoo, Dr. Ven Jyn Kok, and Dr. Sim Ying Ong, thank you for all the advice and encouragements. Abhishek Jhawar, Yang Loong Chang, Ying Hua Tan, Sue Han Lee, Chee Kheng Ch'ng, and Jia Huei Tan, the stimulating discussions, and the joy and laughter we shared throughout our struggles will never be forgoten.

To my family and friends who have always been my pillars of support. No amount of thank yous will be able to show the amount of gratitude I have for the understanding and kindness that I had received. You have given me the strength to presevere until the very end of this undertaking.

An honorable mention goes to Dr. Rodney Tan, who has introduced this path to me, given me continuous moral support, and had faith in me when I had none in myself. You have shown me the beginning, I hope that my success will make you proud.

Last but not least, to everyone whom I have had the fortune to cross paths with these few years. No matter how fleeting our encounters may be, I am grateful for the wisdom and knowledge you have imparted on me. I offer you my blessings and wish you well in your future endeavors.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

Low-light environment is an integral part of everyday activities. As day change to night, the amount of available light decreases, causing the surroundings to be increasingly dark, and subsequently affecting a person's abilities to perform even menial tasks due to a lack of visibility. As studied by Pedersen & Johansson (2016), surrounding illumination affects even the simple action of walking where the reduction of light deteriorates walking quality, whereas Fotios et al. (2015) analyzed that lighting is a key factor of pedestrian reassurance.

On more severe circumstances, low-light can be a cause of accidents and even criminal activity with dire consequences as shown in Fig. 1.1[1]. To illustrate, Calabrese et al. (2017) found that nighttime work are more hazardous for railroad workers with links to darkness as a possible cause, and Pour-Rouholamin & Zhou (2016) identified the time 8 p.m. to 5.59 a.m. and darkness are associated with more severe injuries for pedestrians involved with road accidents in the state of Illinois, United States. Anarkooli & Hosseinlou (2016) investigated the effect of lighting conditions on crash severity which found that dark environments impacts crashes of not only between moving vehicles but also between a moving vehicle and fixed objects due to low visibility, likewise data collected by Khalilikhah & Heaslip (2017) showed that animal-vehicle collisions are higher at night than daytime.

Similarly, research on crime patterns such as those conducted by Tompson & Bowers (2013) found that darkness is significantly associated with increase of street robberies in

[1]Sources:
(a)http://www.dailymail.co.uk/news/article-4084276/BMW-driver-36-dies-horror-crash-car-left-road-careered-gardens-leaving-looking-like-bomb-gone-off.html
(b)http://www.straitstimes.com/asia/se-asia/car-hits-teen-cyclists-in-jb-8-die-and-8-hurt
(c)https://www.washingtonpost.com/news/local/wp/2017/02/25/police-investigating-homicide-in-prince-georges-county/?utm_term=.f4861e0c72f7
(d)https://www.nytimes.com/2017/08/19/us/police-shooting-florida-kissimmee.html

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

Figure 1.1: (a) A car moved off the road and crashed through several front gardens of a neighborhood in Walsall, England on the night of January 2, 2017 at 10.20 p.m., killing the driver, (b) Road accident in Johor Baru, Malaysia on February 19, 2017 where eight teen cyclists were killed around 3.30 a.m., (c) Police investigating a homicide that occurred in Prince George's County, United States on February 25, 2017 after receiving a call at 2.40 a.m., (d) Scene of investigation where two police officers were found shot dead on the night of August 18, 2017 at 9.30 p.m.

London and Glasglow, whereas de Melo et al. (2017) analyzed that severe crimes like homicides, rapes, and robbery often occurs around 7 p.m. to 11 p.m. which is towards the night. Additionally, Hanaoka (2016) reported a significantly higher rate of snatch-and-run at nighttime as compared to daytime in the city of Osaka, Japan, and Montoya et al. (2016) found that more than half of the burglaries in their data occurs at night as well.

Currently, the Close-Circuit Televisions (CCTVs) are increasingly deployed throughout cities in the world especially for crime prevention, however studies have shown mixed results in regards to their effectiveness (H. Lim et al. (2016)). It is noted that such CCTVs are merely recording devices that acts as an archive whenever an incident has already transpired, thus could be the reason that they are not significantly effective. For this reason, computer vision research and systems aimed at assisting people in daily activities, as well as improve safety and security could be especially helpful (Leo et al. (2017)). This is because such systems can imbue CCTVs with artificial intelligence for instantaneous response towards accidents or criminal activities, thus provide more effective assistance and protection.

In the perspective of computer vision and image process, images or videos with low contrast and low brightness/illumination are defined as captured in low-light conditions

and are challenging visual data to work with. However, in terms of research efforts, the low-light domain commonly addresses the image enhancement problem that hardly relates to assistive systems, or night vision surveillance that demands costly hardware that is only practicable for military use at this phase. More relatable subjects, such as object detection, that could help drivers identify "invisible" objects on dark roads or alert authorities of criminal presence in dark alleys, are seldom given attention in low-light research. Hence, it motivates this research to explore low-light vision and provide a solution that would pave the way for future smart low-light assistive and surveillance systems that would benefit the community.

1.1 Problems

There are several problems to be addressed in this work. Firstly, there is a significant lack of data to facilitate and benchmark research efforts in low-light. Even in conspicuous fields, such as object detection that has achieved significant breakthroughs, they evidently deal with bright images while significantly lacking for low-light. For instance, well known public object datasets like PASCAL Visual Object Classes (PASCAL VOC) (Everingham et al. (2010)), ImageNet (Russakovsky et al. (2015)), and Microsoft Common Objects in Context (MSCOCO) (Lin et al. (2014)), have played an integral role in the advancements as they have provided large scale data for many to work on and as challenges that promote progress in object detection and recognition, as summarized in Table 1.1.

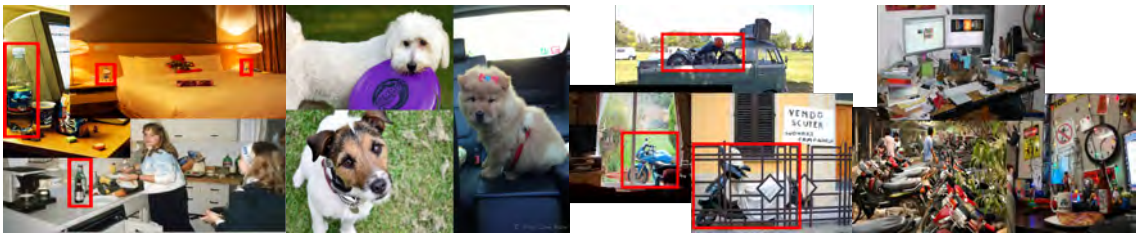
The PASCAL VOC were one of the earlier object datasets with comparatively large amounts of images at that time, consisting many variations that could represent realistic environments during a time where object datasets suffer from simplicity and bias (Torralba & Efros (2011)). Since the launch of the dataset in 2006, it has facilitated the development of many handcrafted approaches for object centric applications (Felzenszwalb

Table 1.1: Summary of renowned public object datasets.

Dataset	PASCAL VOC	ImageNet	MSCOCO
Years active	2005 - 2012	2010 - present	2014 - present
Total classes	20	1,000	80
Annotation	Image class, bounding box, segmentation	Image class, bounding box, SIFT features	Bounding box, segmentation, captions, people keypoints
Total images	26,305	14,197,122	300,000
Approximate low-light images	0.23%	0.03%	1.34%

et al. (2008)). In 2010, the rise of internet data mining has led to the collection of even larger scale data, prominently ImageNet that led to the breakthrough of deep learning using Convolutional Neural Network (CNN) (Krizhevsky et al. (2012)), and subsequently spark a whole new generation of deep learning works in computer vision and machine learning domains. While datasets continue to grow in numbers, a new challenge arises in the form of data annotation because it is difficult for the human annotators to cope with the sheer numbers. Then enters MSCOCO in 2014, though not as large in numbers as the ImageNet, it brings to the table comprehensive annotation covering a variety of tasks which includes recognition, segmentation, and captioning. While the progress brought by these datasets cannot be denied, there is a glaringly obvious lapse, that is, less than 2% of the images provided by these influential datasets are captured in low-light. Moreover, there are no other publicly available datasets that specifically provide low-light images for object focused works to the best of my knowledge.

Subsequently, this led to the second problem, which is an insufficient understanding of the low-light phenomenon in computer vision, especially application based studies. Considering very early computer vision works, such as well-known feature extractors (Dalal & Triggs (2005); Lowe (2004)), had already strove for illumination invariance in their designs, low-light has been treated as an auxiliary element to other tasks. Consequently, state-of-the-art works, both past and present (Felzenszwalb et al. (2008); He et al.



(a) Scale (b) Intra-class variation (c) Occlusion (d) Clutter

Figure 1.2: Example of typical challenges faced in popular object datasets. (a) Scale: Object scale within different images can vary from 90% to 10% of the image size; (b)Intra-class variation: Objects from a single class but having somewhat different appearances with one another; (c) Occlusion: Objects are blocked or only shown partially in the image; (d) Clutter: Highly complex image containing objects of labeled and unlabeled classes. (Source: MSCOCO dataset (Lin et al. (2014)))

(2016); Krizhevsky et al. (2012); Simonyan & Zisserman (2014); J. Wang et al. (2010)), have been designed in such a way to handle comparatively minuscule illumination variations (i.e. shadows) instead of full-fledged low-light conditions (e.g. nighttime). Though not completely devoid of low-light samples in the experiments, they were also scarcely analyzed in such works in favor of other challenges like scale, intra-class variation, occlusion, and clutter as shown in Figure 1.2.

These two problems inadvertently influenced low-light and application based researches to be on two different spectrums. Studies related to low-light itself commonly work on one of two directions, either night vision or image enhancement. Night vision is closely related to surveillance applications (Davis & Keck (2005); J. Dong et al. (2007); Elguebaly & Bouguila (2013); Kang et al. (2014); Qi et al. (2014); Zhao et al. (2015)), however, it hinges on sophisticated hardware that are more suited for military use instead of largescale commercial deployment. On the other hand, low-light image enhancement has been focused on improving visual quality (X. Fu, Zeng, Huang, Liao, et al. (2016); X. Fu, Zeng, Huang, Zhang, & Ding (2016); Guo et al. (2017); L. Li et al. (2015)) without substantial evaluations to show their value for applications like object detection or face recognition. Figure 1.3 shows the common pipelines of these two fields where they both have their own objectives and do not merge into a unified system. While these are

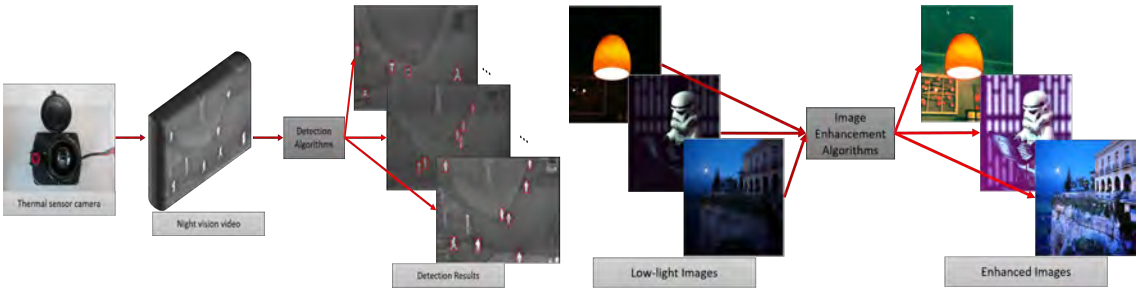


Figure 1.3: Common pipelines of low-light work. (a) Low-light object/person detection using night vision cameras where thermal or near infrared cameras are used to capture a surveillance video, and detection algorithms are built to function based on the characteristics of the videos/frames; (b) Low-light image enhancement where the images are enhanced for better content visibility and aesthetic quality. Both areas of research work on the low-light domain but have different aims.

worthwhile studies to be explored on itself, current works did not show the potential and contribution to the development of a practical, effective and well-rounded intelligent vision system. For that reason, the third problem present is this gap between enhancement and application, in particular, there is a lack of a framework that consolidates these two aspects and improves both the practical performance of applications and the visibility of contents in low-light images.

1.2 Objectives

This work has three objectives to address each of the problems stated in Section 1.1. The first is the collection of a low-light image dataset so as to have a standard benchmarking data to kick start the study. This dataset will not only be used to facilitate the next two objectives of this work, but can also serve as a go-to data for the advancement of low-light research in general.

The second objective is to conduct a comprehensive analysis using the collected data to gain an understanding of the low-light phenomenon, not only from the perspective of low-level vision, but also in the context of applications, such as object detection. This is to gain specific insights that is lacking in the current literature, particularly on the

characteristics and effects which low illumination imparts on computer vision which are commonly simplified as “illumination variation”.

Finally, using the data and understanding gained to develop a computer vision solution that would be the groundwork for the advancement of intelligent vision systems towards low-light functionality. Particularly to bridge the gap between low-light enhancement with object detection, by proposing an image enhancement framework that emphasizes on information retrieval while maintaining fair visual quality. In light of this distinctive objective, new evaluation metrics are to be proposed as well for the assessment of the framework in retrieving informative details as opposed to the image quality centric metrics routinely used in the field.

1.3 Contributions

The contribution presented in this thesis is threefold, to achieve each of the aforementioned objectives in Section 1.2. Foremost, a low-light image only dataset, named the Exclusively Dark (ExDark) dataset is proposed containing 7,363 low-light images from very low-light environments to twilight, with 12 object classes annotated on both image class level and local object bounding boxes, as shown in Figure 1.4. At the time of writing this thesis, this is the largest low-light image dataset with object annotation to-date.

Secondly, an image pixel intensity analysis as well as a feature-based analysis were studied on these low-light images. Specifically, by studying the intensity histograms of low-light images and also the local intensity changes where it is found that low-light images consist of global illumination variation between different low-light images and local illumination variation where intensities vary with respect to light sources either capture within or outside the image. As for feature-based analysis, object proposal methods were used to gage the performance of common hand-crafted features when applied on low-light, while features learned by CNN were studied to gain insights on how machine

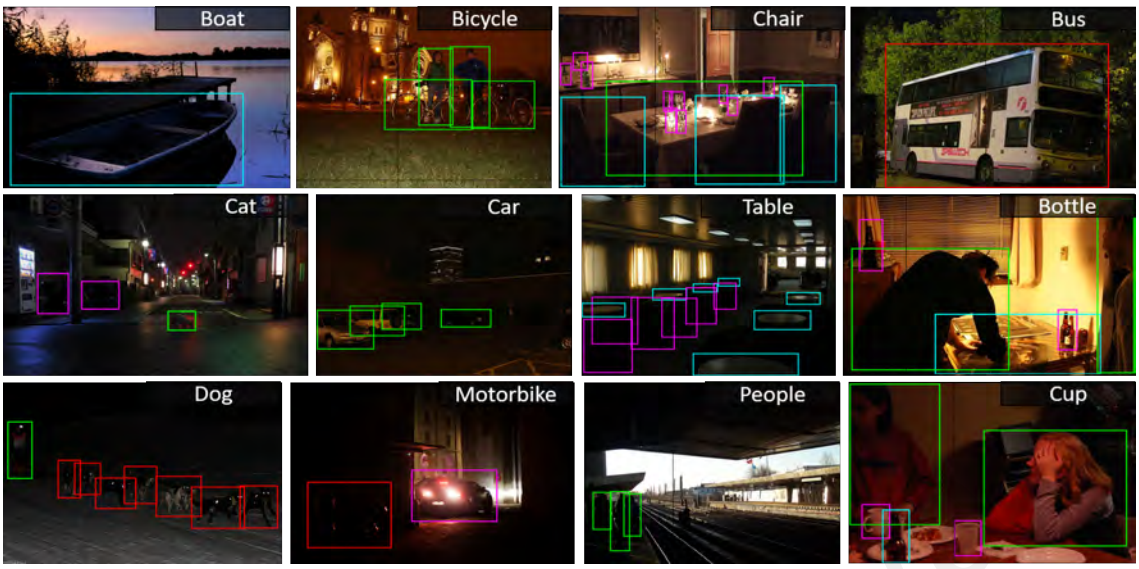


Figure 1.4: Examples of images of every class with image and object level annotation from the ExDark dataset.

perceives and deal with low-light as opposed to vision with sufficient illumination. Not only have the analyses found that conventionally designed features are inadequate for low-light images, but the features learned by machines showed that the illumination variations of low-light essentially require features that are distinct from bright images. This discovery prompts a re-evaluation of the established perception and handling of low-light.

Lastly, based on the understanding that low-light consists of global and local illumination variations, the Gaussian Process ($\mathcal{GP}$) regression (Williams & Rasmussen (2006)) is proposed for low-light image contrast enhancement due to its sophistication and robustness in modeling localized data, supported by an intermediate CNN model to produce the necessary priors learned from globally generalized large data. Figure 1.5 illustrates the overall framework proposed. The standout advantage of this solution is its ability to retrieve informative details (i.e. features) in its enhancements while maintaining fair visual quality, as affirmed by the conventional image quality metric, the PSNR), and new evaluation metrics focused on features, namely, local features matching and intensity histogram distance.

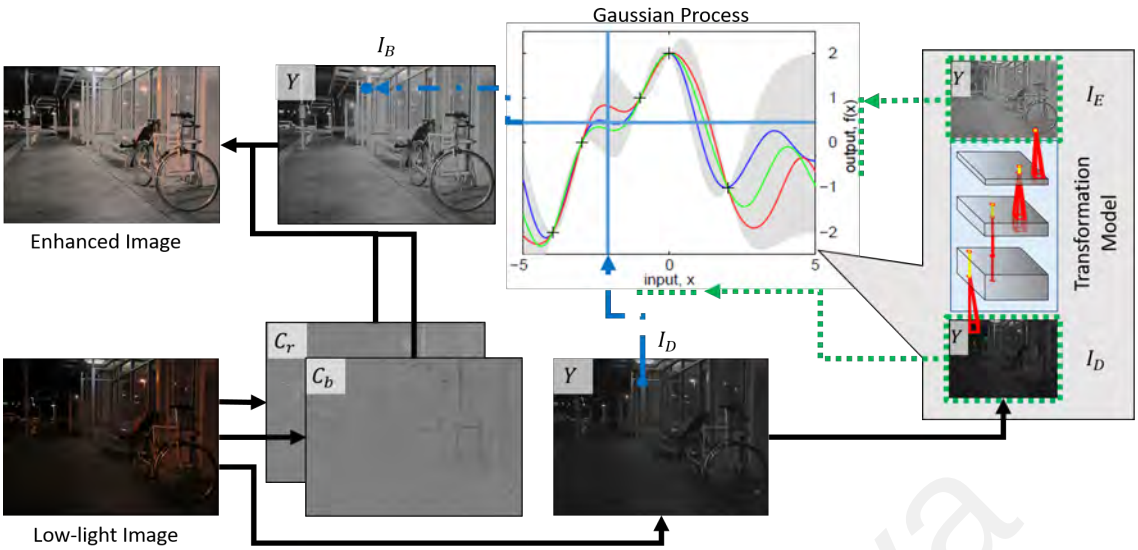


Figure 1.5: Overall framework of the proposed low-light image enhancement using $\mathcal{GP}$ with CNN data.

1.4 Outline

This chapter details the overall motivation, problems, objectives, and contributions of this thesis. Brief outline of the remaining chapters are as follows:

Chapter 2 reviews the current literatures related to this work. In particular, the existing object image datasets and their lack of low-light images, followed by the common datasets used in low-light research such as the night vision camera captured data. Additionally, existing works on low-light image enhancement, their shortcomings and inadequacy in supporting object detection are also discussed here.

Chapter 3 introduces the low-light image dataset, the ExDark dataset that would be used throughout the study of this thesis. It includes discussions on the evolution of the object datasets used in the computer vision research community throughout the years and the manner in which it has inspired this work and the proposal of this dataset that is yet unheard-of in the community to the best of my knowledge. The dataset content and statistics are detailed in this chapter.

Chapter 4 describes the analysis of low-light images. Starting with low-level analysis that looks into the variations of pixel intensities in both the global and local scale. This

is followed by the high-level analysis, with the implementation of conventionally hand-crafted features like edges, gradients, and superpixels through object proposal methods and the comparison of their behaviors on both low-light and bright images are made clear. Furthermore, details in regards to the employment of deep learning, namely CNN to learn features, the scrutiny of the said features, and also the insights gain are explained here as well.

Chapter 5 proposes the low-light image enhancement framework, namely Gaussian Process with CNN data. This chapter contains a theoretical overview of the $\mathcal{GP}$ algorithm, the justification of its use for low-light image enhancement, particularly for the objective of information retrieval, as well as the rationale of the CNN data. Comprehensive experiments were done using real and synthetic low-light images, with evaluations using both the PSNR and the new metrics. Both qualitative and quantitative results in comparison to current state-of-the-art methods are detailed, where the proposed method shows promising results.

Chapter 6 concludes the work and findings obtained from this study with current limitations and suggestions for the prospective future developments.

CHAPTER 2: LITERATURE REVIEW

This chapter reviews literature related to the work of this thesis. It is divided into two main categories, starting with the review on dataset related literatures, which are subdivided into object datasets, low-light data, and alternative datasets. This is then followed by review on low-light research works, particularly low-light image enhancement.

2.1 Datasets

Datasets are an important element in benchmarking all research works. For this work, a low-light object dataset is required, however, it was found that current publicly available object datasets do not have sufficient low-light data for adequate benchmarking whereas data used in low-light research are neither suitable nor sufficient for this study.

2.1.1 Object Datasets

There has been three major publicly available object datasets that are renown to every researcher in this field.

PASCAL VOC: The PASCAL VOC (Everingham et al. (2010)) object dataset grew from 2005 till 2012, with annual challenges that encouraged researchers to develop ever improving algorithms to outdo one another in the spirit of progress. It began with only 4 object classes and 3,787 images sourced from existing datasets. Initially containing simple object images, it has been continuously improved with more challenging images, and additional annotations. The last update to the dataset in 2012 puts the cumulative total at 26,305 images with 20 object classes, including annotations for object region of interest and segmentations.

ImageNet: ImageNet (Russakovsky et al. (2015)) was opened to public in 2010 as the largest object image dataset, and gained great interest from the community espe-

cially in 2012 where its database of over 1 million images and 1,000 image level object classes has allowed CNNs to excel and set a new benchmark in object image classification (Krizhevsky et al. (2012)). The data provided are very challenging, where each of the image is categorized into one of the object classes as long as there are instances of the object, regardless if the objects are occluded or if the image contains other objects. Since then, ImageNet has become the de facto dataset for object image works, either as the main benchmark (Krizhevsky et al. (2012)) or as fundamental data for transfer learning (Donahue et al. (2014); Lee et al. (2017); Tong et al. (2016)). In 2017, the dataset has reach new heights with more than 14 million images, and 1,000 classes of which 200 of them has bounding box annotation for object detection tasks.

MSCOCO: The latest of notable object datasets is the MSCOCO (Lin et al. (2014)), released in 2014. The quantity of images provided are not up to that of ImageNet, though its advantage is in the completeness of the image annotations. Providing more than 300 thousand images in 2017, there are 80 object classes annotated from bounding box for detection, to pixel level for segmentation tasks, as well as captions for description of each image. Similar to ImageNet, the content of the images are highly challenging where even a small instance of an object's part is annotated.

Though challenging and large, the number of low-light images in these 3 datasets are considerably small, as shown in Table 2.1. This brings about a difficulty in understanding the effects of low-light and insufficient as a benchmark.

2.1.2 Low-light Data

On the other hand, datasets used in low-light research for the most part are different from typical object detection datasets. One of which is used in surveillance that can be categorized as detection tasks, but the data used differ greatly from typical object detection due to the use of different types of cameras. The other is in enhancement, where

Table 2.1: Approximate number of low-light images in PASCAL VOC, ImageNet, and MSCOCO. Even as the years progress, there is no significant increase of low-light images in either of the datasets.

Dataset		Total image	Low-light image
MSCOCO	Training	82,783	149 (0.18%)
	Validation	40,504	163 (0.4%)
	Testing 2014 (No annotation)	40,775	138 (0.34%)
	Testing 2015 (No annotation)	81,434	115 (0.14%)
	Total	245,496	**565 (0.23%)**
ImageNet	Training 2012	1,300,000	255 (0.02%)
	Validation 2012	50,000	38 (0.08%)
	Testing 2012	100,000	51 (0.05%)
	Validation 2013	4,599	12 (0.26%)
	Testing 2013	9,251	22 (0.23%)
	Training 2014	60,658	72 (0.12%)
	Total	1,524,508	**450 (0.03%)**
PASCAL VOC	2007	9,936	123 (1.24%)
	2008	4,340	72 (1.66%)
	2009	2,722	43 (1.58%)
	2010	3,503	50 (1.43%)
	2011	3,640	48 (1.32%)
	2012	2,164	17 (0.79%)
	Total	26,305	**353 (1.34%)**

algorithms are proposed to improve the visibility of the contents in low-light images.

Low-light surveillance: Thermal and near infrared cameras are generally used to counter limited light for surveillance operations at night. Surveillance works commonly focus on face recognition (Kang et al. (2014); S. Z. Li et al. (2007)) and pedestrian detection (Davis & Keck (2005); J. Dong et al. (2007); Qi et al. (2014); Zhao et al. (2015)). Datasets in this field are usually acquired using sophisticated hardware such as thermal sensors and Infrared (IR) cameras, as shown in Table 2.2. These equipment captures visual contents beyond the visible spectrum of humans, which overcomes the lack of light encountered in nighttime and low-light environments.

Thermal imaging, or Far-Infrared (FIR), employs passive sensors to capture infrared radiation emitted by objects from a scene. This radiation emission is associated with the temperature where higher temperatures often corresponds to higher emissivity. Based on this property, the visual data captured can clearly show objects (such as people having

Table 2.2: Hardware and setups implemented to obtain data for research domains beyond visible spectrum .

Domain	Literature	Hardware	Setup
Pedestrian detection	Davis & Keck (2005)	Raytheon 300D thermal sensor	Mounted on rooftop of 8-story building
	Davis & Sharma (2007)	Raytheon PalmIR 250D; Sony TRV87 Handycam	Mounted adjacently, approximately 3 stories above ground
	Bilodeau et al. (2014)	FLIR Thermovision A40M; Sony XCD-710CR	Indoor with fixed background and temperature
	Olmeda et al. (2013)	Indigo Omage Imager	Mounted on vehicle exterior to avoid infrared filtering
Face recognition	S. Z. Li et al. (2007)	Specially designed NIR hardware using active lights in the Near-Infrared (NIR) spectrum	
	Kang et al. (2014)	Canon 600D DSLR, RayMax 300 NIR illuminator	Stationary camera and illuminator positioning with specified emission direction
Object detection	Z. Wu et al. (2014)	FLIR SC8000	-

warm bodies) from the background irrespective of light level, which is ideal for surveillance, as seen in the top row of Figure 2.1. However, as shown in Table 2.2, these data are either captured by advanced cameras or specially designed acquisition setup. Consequently, such setup is not as commonplace as visible light cameras.

On the other hand, NIR uses the infrared spectrum from thermal sensors and less affected by temperature. As shown in the bottom row of Figure 2.1, the appearance of the captured contents capture by NIR are more similar to visible light images. However, as the name suggests, the operational distance is very much shorter, as illustrated in Figure 2.1 where it is difficult to see the person at far distance, unlike thermal imaging. While this limitation does not interfere with indoor and short distance monitoring, it greatly obstructs its employability in outdoor surveillance.

Low-light enhancement: To the best of my knowledge, current low-light image

Figure 2.1: Top row: Examples of thermal images and their corresponding images captured in visible light. Bottom row: Example images of a person at short (1m) and long (150m) distances captured using NIR in low-light and using digital camera in bright environment.

enhancement works do not have a go-to benchmarking dataset. As a result, the data used are varied and nonstandard, such as datasets catered for other type of enhancement works. The IVC database (Le Callet & Autrusseau (2005)) is a quality metric data that is commonly used by works involving quality assessment. Therefore, works like those done by J. Lim et al. (2015); Lore et al. (2017), used this dataset by synthetically darkening the images to simulate low-light conditions as shown in Figure 2.2; at the same time, the original bright images are used as the groundtruth for evaluating the enhancement results. While such synthetic darkening is practical for evaluation, there is only 235 images in the IVC dataset, an awfully small amount when compared to datasets in the object detection domain.

This is similarly seen in the data of real low-light images. S. Wang et al. (2013) has proposed the Non-Uniform Illumination (NUI) dataset of images collected from online websites with only 156 images, whereas other works like those done by X. Fu, Zeng, Huang, Liao, et al. (2016); Guo et al. (2017); Huang et al. (2013); L. Li et al. (2015) capture or download low-light images in an ad-hoc manner. Figure 2.2 shows examples of these images. Not only were these data highly inconsistent, the images are catered for

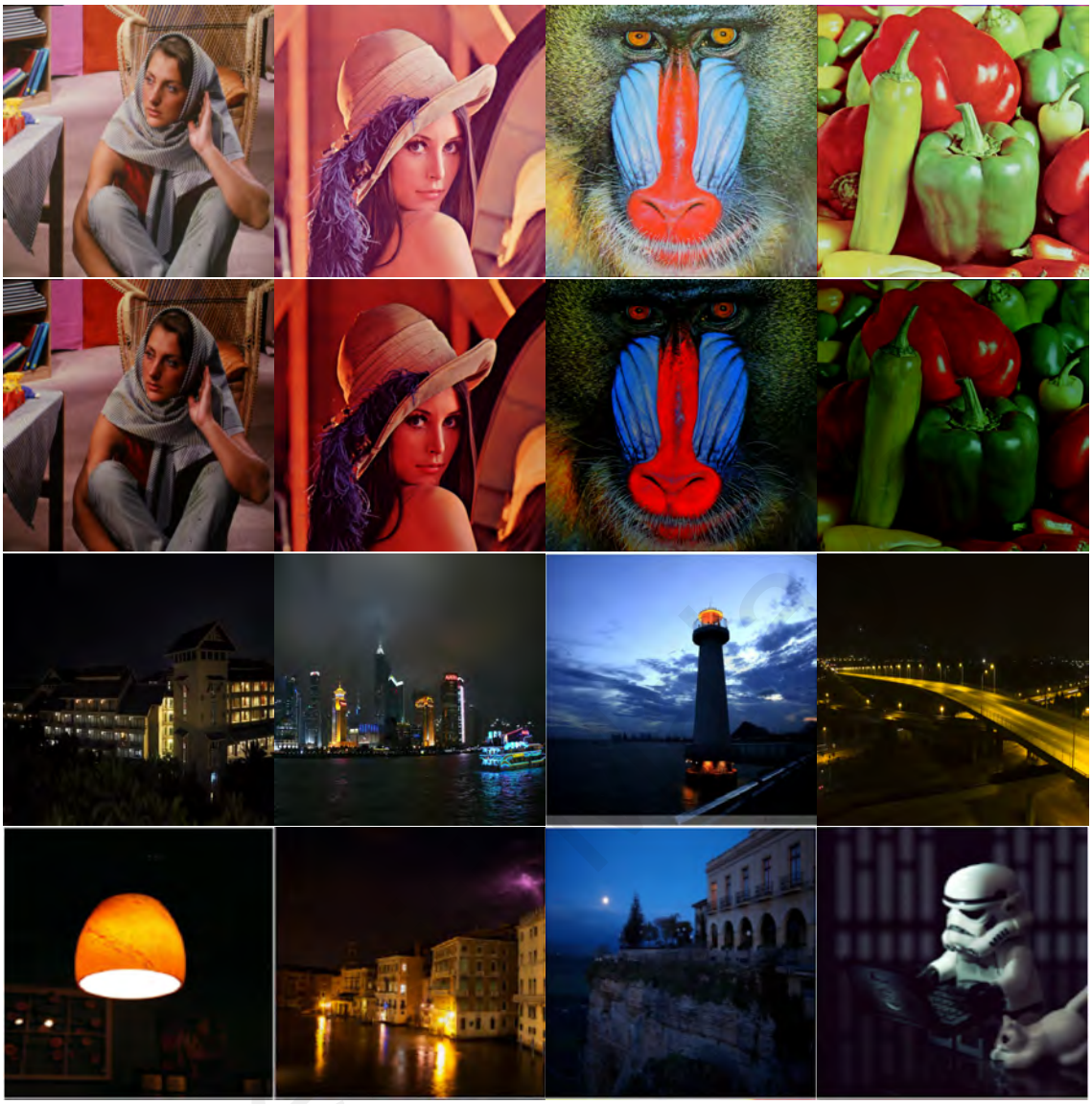


Figure 2.2: Examples of images used in low-light image enhancement works. First row: IVC dataset; Second row: Example of synthetically darkened images of IVC dataset; Third row: NUI dataset; Third row: Images used by Guo et al. (2017).

quality assessments instead of object based evaluation, i.e. low-light images of scenery instead of images containing everyday objects. Hence, these current data are neither adequate nor enough for the development and gaging of object detection in low-light.

2.1.3 Alternative Datasets

Alternatively, there are public datasets available from related research domains, namely the Phos (Vonikakis et al. (2013)), DaLI (Simo-Serra et al. (2015)), Webcam (Verdie et al. (2015)), and ALCN-2D (Rad et al. (2017)) datasets. Such datasets provide images for

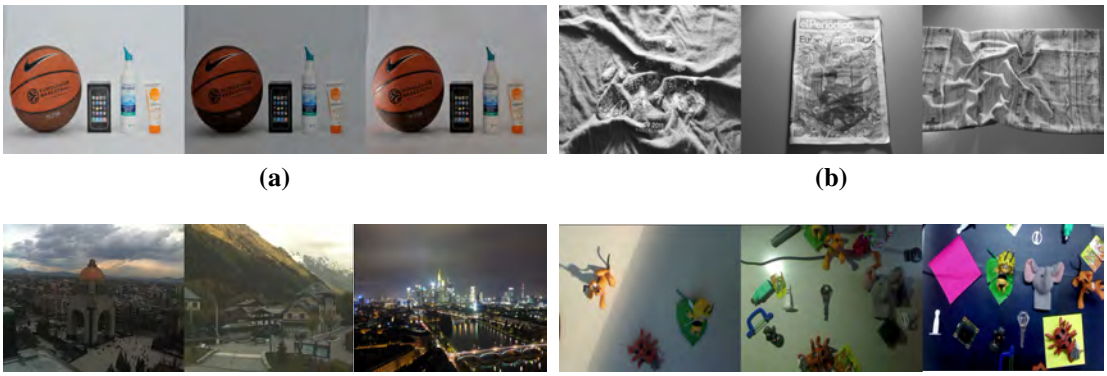


Figure 2.3: Examples of images from alternative datasets, (a) Phos, (b) DALI, (c) Webcam, and (d) ALCN-2D.

illumination variation research works with fundamentally different motivations, where the images are either captured under controlled environments (Phos, DALI, ALCN-2D) or scenery images (Webcam), as shown in Fig. 2.3.

The Phos dataset is a database of images captured under different illumination conditions. It provides a baseline exposure image as a recommended reference, captured under uniform illumination achieved by multiple diffusive light sources distributed around the objects and standard exposure. The dataset contains both uniform and non-uniform illumination images, where various levels of uniform illumination images are captured by reducing the intensity of the diffusive lights, while the non-uniform illumination images are captured at reduced diffusive light strength with a strong directional light source on the left.

Whereas, the DaLI and Webcam datasets are provided for feature descriptors and keypoints evaluation works. The DaLI contains images with uneven illumination caused by deformations like wrinkles on a t-shirt or paper, while the Webcam consists scenery images of up to 6 locations. Lastly, the ALCN-2D provides data for object detection under challenging lighting conditions with background clutter. However, the objects in question are not common objects like those provided by MSCOCO but are only 3 small objects made of different materials.

Although these public datasets work on lighting research, they are fundamentally different from the target data of this research which focuses on complex low-light images of real low-light environments instead of controlled laboratory settings.

Tables 2.3 and 2.4 show a summary of the aforementioned publicly available datasets related to this research work. On one hand, application based research data either lacks low-light data or is compensated by using night vision hardware. On the other hand, the data used in low-light related works are either too small, or captured under constrained environments that do not represent the challenges of the real world. Therefore, their unsuitability to facilitate a study on computer vision applications in real low-light environments as well as a lack of a standard repository calls for a collection of a new database.

2.2 Low-light Image Enhancement

From the investigation on datasets, it is noted that research works in computer vision tackle low-light problems from one of two angles, (1) through the use of hardware supports such as cameras equipped with thermal sensors and infrared cameras; or (2) using enhancement algorithms on low-light images or videos (X. Dong et al. (2011); H. Fu et al. (2012); X. Fu, Zeng, Huang, Liao, et al. (2016); X. Fu, Zeng, Huang, Zhang, & Ding (2016); Guo et al. (2017); Huang et al. (2013); L. Li et al. (2015); J. Lim et al. (2015); Łoza et al. (2013); X. Wu (2011); X. Zhang et al. (2012)). As discussed in Section 2.1.2, hardware requirement are costly and to-date is uncommon and impractical for large-scale deployment. Whereas the latter, namely low-light image enhancement are generally applied for image quality improvement, however these algorithms mainly deal with visible light images which are significantly more easily found and produced. Thus, low-light image enhancement holds more potential to support the development of intelligent computer vision applications. The various proposed works of low-light image enhancement can be categorized into three categories, statistical model and manipulation, transforma-

Table 2.3: Comparison of existing related datasets (object detection and pedestrian detection) that is publicly available.

Domain	Literature	Imaging type	Amount	Information
Object detection	PASCAL VOC (Everingham et al. (2010))	Visible light	26,305	Object images from unconstrained background and locations, majority captured in daylight.
	ImageNet (Russakovsky et al. (2015))	Visible light	14,197,122	Object images from unconstrained background and locations, majority captured in daylight.
	MSCOCO (Lin et al. (2014))	Visible light	300,000	Object images from unconstrained background and locations, majority captured in daylight.
	Thermal Infrared Video Benchmark for Visual Analysis (Z. Wu et al. (2014))	Thermal	63,782	Frames from 5 outdoor and 2 indoor scenes capturing multiple moving objects such as pedestrians, vehicles and flying animals.
Pedestrian detection	OSU Thermal Pedestrian Database (Davis & Keck (2005))	Thermal	284	Fixed view of walkway and street with pedestrians, captured in the morning and afternoon, with some rainy weather condition.
	OSU Color-Thermal Database (Davis & Sharma (2007))	Thermal and visible light	17,082	Frames from 6 video sequences, fixed view of two locations with pedestrians, captured at different times-of-day.
	LSI Far Infrared Pedestrian Dataset (Olmeda et al. (2013))	Thermal	81,592	Outdoor urban scenario from moving vehicle with pedestrians.
	Pedestrian Infrared/visible Stereo Video Dataset (Bilodeau et al. (2014))	Thermal and visible light	5,390	Indoor close range video frames of pedestrians.

Table 2.4: Comparison of existing related datasets (face recognition, low-light enhancement, and illumination research) that is publicly available.

Domain	Literature	Imaging type	Amount	Information
Face recognition	CBSR NIR Face Dataset S. Z. Li et al. (2007)	NIR	3,940	Face images captured in indoor environment with varying light conditions.
	LDHF Database (Kang et al. (2014))	NIR and visible light	800	Fixed indoor and outdoor images of frontal face view, varying distance, captured at daytime and nighttime
Low-light enhancement	IVC database (Le Callet & Autrusseau (2005))	Visible light	235	10 bright images of various objects, remaining images generated by different distortions. Used in low-light research by synthetic darkening.
	NUI dataset (S. Wang et al. (2013))	Visible light	156	Scenic images captured manually and downloaded online.
Illumination research	Phos (Vonikakis et al. (2013))	Visible light	210	Images of specific objects arranged in constrained lab environment, captured in controlled uniform and non-uniform lighting conditions.
	DaLI dataset (Simo-Serra et al. (2015))	Visible light	192	12 objects captured in lab under different illumination and deformation.
	Webcam dataset (Verdie et al. (2015))	Visible light	850	Scenery images of 6 locations captured using outdoor webcam at different times of day and seasons.
	ALCN-2D dataset (Rad et al. (2017))	Visible light	3,630	Images of 3 objects made from different materials captured in a lab under varying illumination, lighting color, and background distractors.

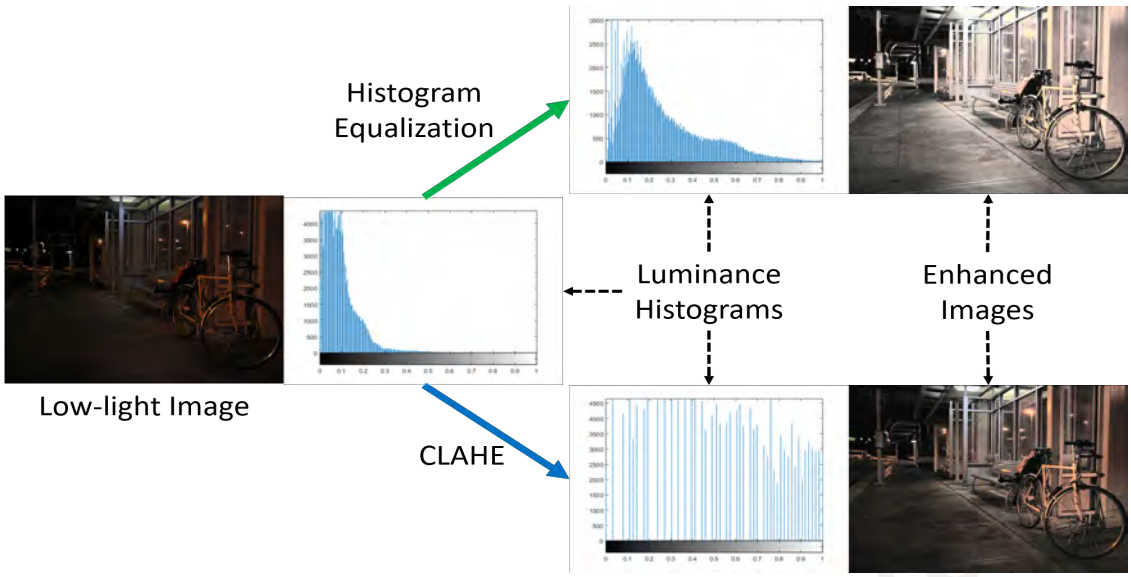


Figure 2.4: Low-light image enhancement by histogram equalization and CLAHE. The luminance histogram of a low-light image is concentrated towards the darker regions. These methods pushes the luminance counts to a wider range for brighter values.

tion model, and Retinex model.

Statistical model and manipulation: This category of approach manipulates the distributions of low-light images, either intensities or high frequency coefficients, to improve the image contrast and brightness (Huang et al. (2013); J. Lim et al. (2015); Łoza et al. (2013)). This category includes the earliest approach for contrast enhancement, the histogram equalization and its variants (Huang et al. (2013); Kaur et al. (2011)). Figure 2.4 shows example of the luminance histogram enhanced by histogram equalization and Contrast Limited Adaptive Histogram Equalization (CLAHE) (Zuiderveld (1994)), the most common methods in enhancing image contrast, where it can be seen that the luminance histograms are notably modified. Later works incorporate more sophisticated mechanisms, for example, J. Lim et al. (2015) used noise pixels to guide a selective histogram equalization scheme for simultaneous denoising and contrast enhancement. However, such approaches commonly expand or stretch the concentrated intensity histograms of low-light images, hence loses contextual information and prone to under enhancement or over enhancement.

Transformation Model: This approach (H. Fu et al. (2012); Lore et al. (2017); X. Wu (2011)) uses parameterized functions or trained models to perform transformation mapping from low-light image space to bright image space, meanwhile preserving the contextual information. Following the achievement of deep learning approaches in computer vision, Lore et al. (2017) trained a deep auto-encoder model from large amounts of data to obtain a general mapping model for contrast enhancement. The enhancement of these methods are defined by inferring parameters that would produce satisfactory outcomes for a variety of conditions. Nonetheless, they rely on generalization from a database of images, and concentrate on contrast enhancement without investigating the effects brought upon by local illumination variations, thus led to non-optimal enhancement.

Retinex model: Based on the Retinex theory by Land et al. (1977), it takes into consideration both the contextual information and light intensity of an image. The main assumption is that a color image can be decomposed to reflectance and illumination components to represent the aforementioned elements respectively. By manipulating the illumination component and merging with the reflectance, various methods (X. Fu, Zeng, Huang, Liao, et al. (2016); X. Fu, Zeng, Huang, Zhang, & Ding (2016); Guo et al. (2017)) had shown impressive results. X. Fu, Zeng, Huang, Liao, et al. (2016) used a fusion-based method where the illumination is enhanced using multiple methods separately before fusing them in multiple scales with heuristic weights for a final improved illumination to form the enhanced image., whereas X. Fu, Zeng, Huang, Zhang, & Ding (2016) focused on solving the reflectance and illumination decomposition problem using the proposed weighted variational model where the reflectance gives better details while the illumination is only improved using gamma correction. On the other hand, Guo et al. (2017) targets only the illumination where they estimate the illumination of each pixel individually to form a map, and then further refine it using a structure prior. These methods

have achieved state-of-the-art performance in terms of quality, nonetheless they still suffer from under enhancement, noise amplification, and unrealistic color.

Additionally, X. Dong et al. (2011); L. Li et al. (2015); X. Zhang et al. (2012) have proposed methods that resemble the Retinex theory by implementing the dark channel prior algorithm made for image dehazing (He et al. (2011)). This approach is mainly sparked by the observation where inverted low-light images exhibit similar characteristics to images captured in hazy weather. The work by L. Li et al. (2015) has showed state-of-the-art results by combining an adaptive dehazing algorithm with superpixel denoising. While the results are promising, under enhancement is still a prevailing problem that calls for improvement. It is also noted that this approach has inspired works using the Retinex model (X. Fu, Zeng, Huang, Liao, et al. (2016); Guo et al. (2017)) in their illumination map estimation.

A comparative summary of the existing methods are shown in Table 2.5. The proposed solution in this thesis is intrinsically different from these approaches due to the distinctive motivation in performing low-light contrast enhancement. The $\mathcal{GP}$ is a sophisticated statistical modeling technique that interprets the enhancement operation from a localized distribution of functions that is essentially dissimilar to direct manipulation on histogram of pixels intensities, and estimations based on reference maps. Additionally, the implementation of the CNN as an intermediate transformation model introduces reference data that are optimized globally across large data to build the $\mathcal{GP}$. Hence, the localized model with support from globally optimized data is able to perform optimal enhancements for each low-light image.

Table 2.5: Existing research works on low-light image enhancement.

Category	Literature	Color space	Method
Statistical	Huang et al. (2013)	Value (V) component of HSV	Adaptive gamma correction with weighted distribution modification on the intensity histogram.
	J. Lim et al. (2015)	Grayscale	Histogram equalization on Gaussian approximated distribution of noise-free image pixels.
	Łoza et al. (2013)	Value (V) component of HSV	Dual-Tree Complex Wavelet Transform (DT-CWT) on V, exponential contrast enhancement on high-pass wavelet coefficients and CLAHE on low-pass coefficients.
Transformational	X. Wu (2011)	RGB	Optimal contrast tone mapping by linear programming.
	H. Fu et al. (2012)	RGB	Color estimation model modulated by a single parameter determined by statistical observation on large data and then post processed by sparse coding.
	Lore et al. (2017)	Grayscale	Stacked sparse denoising autoencoder model trained on synthesized low-light images.
Retinex	X. Fu, Zeng, Huang, Liao, et al. (2016)	RGB	Weighted multi-scale fusion of luminance and contrast improved illumination component of the Retinex model.
	X. Fu, Zeng, Huang, Zhang, & Ding (2016)	RGB and HSV	Retinex model decomposition by weighted variational model and illumination enhancement by gamma correction.
	Guo et al. (2017)	RGB	Illumination map estimation and refinement by structure-aware smoothing model.
Others	X. Dong et al. (2011)	RGB	Invert low-light image and apply dark channel prior dehazing algorithm He et al. (2011).
	X. Zhang et al. (2012)	RGB	Enhancement using dehazing algorithm where light transmission is estimated from the image luminance instead of the scene depth as used in the standard dehazing.
	L. Li et al. (2015)	RGB	Dehazing with adaptive weight coefficient for light transmission estimation.

CHAPTER 3: THE EXCLUSIVELY DARK LOW-LIGHT IMAGE DATASET

Dataset is an essential part of computer vision research. It represents the environment and circumstances of the real world which intelligent vision systems would need to address. Thus, a reliable and domain specific dataset is needed for an impartial and valuable research. In this work, the target domain is intelligent low-light vision system, hence, the relevant dataset would be images of objects in low-light environments.

3.1 Progression of Object Datasets

The data used throughout the years of vision research have been continuously evolving to better represent the visual perception of the real world, especially for object based research. Figure 3.1 shows a summary of the progression of object image datasets. In the early days, object datasets are collected in a controlled laboratory environment with fixed background, such as the Columbia Object Image Library (COIL) database (Nayar et al. (1996); Nene et al. (1996)). Then, the Caltech dataset (Fei-Fei et al. (2007); Griffin et al. (2007)) grew the nature of object data to include real world backgrounds as opposed to the plain surroundings. Making use of online search engines and then manual filtering, two Caltech datasets were released, the Caltech-101 and Caltech-256, in the years 2003 and 2005 respectively, containing over 30 thousand images with image level annotation for the object classes. Sometime after, Microsoft Research in Cambridge (MSRC) proposed the MSRC dataset (Winn et al. (2005)), though smaller in size, but it gives a stronger annotation using both localized bounding boxes and segmentation. Similarly, the LabelMe (Russell et al. (2008)) provides a dataset with segmentation groundtruth and notably released an online annotation tool for labeling image databases of computer vision research. Both the MSRC and LabelMe datasets also provide data that are somewhat more complex than the object centric data of the Caltech datasets. Figure 3.2 shows examples of images

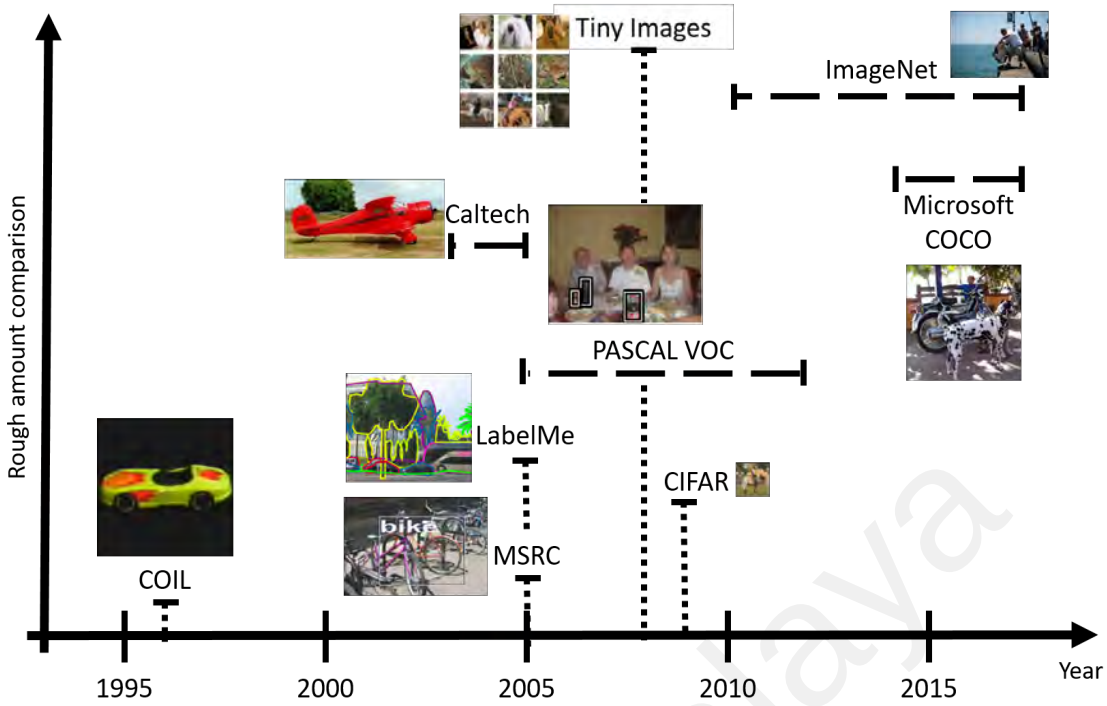
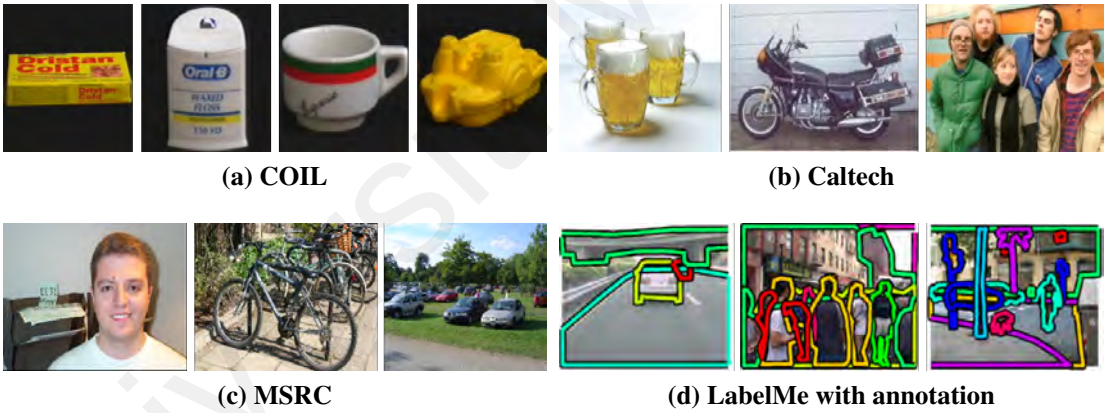


Figure 3.1: Progression of modern object datasets from year 1995 to 2017.



(a) COIL **(b) Caltech**

(c) MSRC **(d) LabelMe with annotation**

Figure 3.2: Examples of images from early object datasets.

from these early object datasets.

After these datasets, the PASCAL VOC ushered in a new trend that not only relies on data itself for advancement. While starting with less than 5 thousand object images with only 4 object classes in 2005, it has continuously grown until 2012 to have over 20 thousand images and 20 object classes with images of objects in varying real world environments, as shown in Fig. 3.3a. More remarkably is the open performance evaluation challenges that enabled researchers in the same field to compare their performances on the same benchmark, which has promoted friendly competition between scientists and led to

significant progress. Unfortunately, the growth of this dataset was cut short by the passing of its main contributor. Though the efforts continues in spirit through other datasets, such as the ImageNet with its ImageNet Large Scale Visual Recognition Challenge (ILSVRC), and the MSCOCO dataset and challenge.

In 2008, the Tiny Images dataset (Torralba et al. (2008)) has amassed close to 80 million images that are downloaded from the internet using several search engines following approximately 75 thousand non-abstract nouns of the WordNet lexical dictionary. However, due to storage constrains at the time, images of this dataset were kept at 32×32 pixels, as shown in Fig. 3.3b. Moreover, the sheer number of data makes it impractical to ensure the quality and annotation using manpower. Thus, in 2009, two subsets were manually extracted from Tiny Images, namely CIFAR-10 and CIFAR-100. Human labelers were assigned to extract 6000 images of 10 object classes for the CIFAR-10, whereas 600 images were extracted of 100 classes exclusive from the CIFAR-10, makes up the CIFAR-100. Thus, the images with strong labeling are lesser in comparison to the amount of data that are actually gathered.

The ImageNet dataset (Russakovsky et al. (2015)) and the ILSVRC started in 2010, but its rise in prominence in 2012 is closely linked to the advancement of Graphics Processing Units (GPUs) and CNN technologies. The ImageNet data are similarly crawled online based on the hierarchy of WordNet. With millions of images crawled, the quality and annotation are done manually to have an average of 1000 images for 1000 categories each and unlike Tiny Images, the images are not resized to accommodate for storage limitations. The image categories provided are much more finegrained than the PASCAL VOC or any dataset before that, such as dogs labeled following the breed instead of one high level category of "dog". The dataset now has over 14 million images and 1 million object bounding box annotations as seen in Fig. 3.3c.

The immense amount images collected by Tiny Images and ImageNet are thanks to

(a) PASCAL VOC

(b) Tiny Images

(c) ImageNet

(d) MSCOCO

Figure 3.3: Examples of images from large scale object datasets.

the rapid expansion of the internet and social media. While it has somewhat solved the problem of data quantity, another obstacle arises in terms of reliable annotation. This aspect is particularly important to the presently popular deep learning methods that needs labeled data for training. Hence, the MSCOCO (Lin et al. (2014)), which is one of the latest and arguably one of the most popular object image dataset now, focuses on providing well labeled data for a wide range of object related tasks. Though only having 330 thousand images, more than 200 thousand are labeled with 80 object categories and up to 1.5 million object instances. The object instances include both bounding box coordinates and superpixel segmentation. Moreover, the dataset also has segmentation of up to 91 stuff categories (i.e. non-object classes like grass, wall, and sky) and annotation of 5 captions describing the content for each image. Examples of images from this dataset with their segmentation annotations are shown in Fig. 3.3d.

Based on the progression of the past 20 years, it has become a standard for object

datasets to have either if not all of the following traits:

- **Complex images**: objects are not necessarily central or dominant in the image, or only partially seen.
- **Large quantity**: from thousands to millions of images.
- **Thorough annotation**: at least having image and bounding box annotation of objects.

3.2 The Exclusively Dark

A significant motivation in the effort to introduce a singular low-light image dataset is that there are none available to-date to set the standards for research in this domain. As seen from the progression, large scale object datasets (Everingham et al. (2010); Lin et al. (2014); Russakovsky et al. (2015)) claim data variations and generalization, however they hardly provide enough low-light data, as shown in Table 3.1, to represent the true extend of environments and challenges faced in such conditions despite being an integral element in daily vision. On the other hand, even in low-light image enhancement works, real low-light images were mostly downloaded or captured on an ad-hoc basis (X. Fu, Zeng, Huang, Liao, et al. (2016); Guo et al. (2017); Huang et al. (2013); L. Li et al. (2015)). Hence, the ExDark is proposed in hopes of providing a staple collection of data for benchmarking low-light research works, and bring together different areas of expertise to focus on low-light, for instance, image understanding, image enhancement, object detection, etc.

3.2.1 Data Collection

The ExDark is targeted to be a low-light object image dataset. An image is defined as captured in low-light if it has low or significant variations in illumination that causes an

Table 3.1: Approximate number of low-light images in public object datasets, and the amount in the proposed ExDark dataset.

Dataset	Low-light image
MSCOCO	565 (0.23%)
ImageNet	450 (0.03%)
PASCAL VOC	353 (1.34%)
ExDark	7,363 (100%)

image to have low contrast and brightness as well as low visibility of the image content. 12 object classes are established for this dataset, namely *Bicycle, Boat, Bottle, Bus, Car, Cat, Chair, Cup, Dog, Motorbike, People,* and *Table*. These classes are derived from the 20 classes of common objects from the notable PASCAL VOC dataset and specifically chosen for their relevance in assistive and surveillance operations, for example, the identification of such objects can help a person navigate around them in the dark, or warn a driver of such objects obstructing a dark road. Following the success of past object datasets, the collection and selection of images are aimed to satisfy the 3 common characteristics, image complexity, quantity, and annotation.

As this is the first dataset of its kind, the collection of data is done through meticulous manual selection instead of the automated download that would result in large amounts of noisy images. The images are collected from a variety of sources targeting the specified object classes, which include downloading from internet websites, sub-sampling from existing datasets, extracting frames from movies and videos, and capturing using smart phones.

Most of the low-light images were downloaded from internet websites and search engines, namely *Flickr.com, Photobucket.com, Imgur.com, Deviantart.com, Gettyimages.com,* and *Google Search*. Keywords related to low-light are used to manually search and download images from these websites, such as *dark, low-light, nighttime, twilight, darkness,* and *shadow*. Only images containing the relevant objects from the listed 12 classes are downloaded for the dataset. Additionally, a combination of object class names and low-

light related words are also used for the search to increase the search coverage for more data.

Next, low-light images were sub-sampled from existing datasets, mainly PASCAL VOC, ImageNet, and MSCOCO, where Table 3.1 shows the approximate mount extracted from each of them. In addition, small amounts of images are also taken from other datasets (Philbin et al. (2008); Russell et al. (2008)). In the same manner, only low-light images containing the relevant objects are extracted for the ExDark. Additionally, to increase the variation of the images, frames from low-light scenes from a collection of movies were extracted as well (See Appendix A for list of sources). Lastly, low-light images were captured manually using different models of smart phones (e.g. Apple iPhone 5S, Samsung S7, Huawei P9).

Approximately 7,363 RGB color images were collected and processed to remove digitally placed watermarks, either by cropping the image or blending it to the background using appropriate coloring. No further processes were applied to the dataset images, thus the dataset contains images of varying quality and sizes from 200×200 pixels to 4000×3300 pixels. Furthermore, all the images are either *jpeg* or *png* formatted, as these two are the most common types of image compression.

3.2.2 Object Annotations

The collected data is manually annotated on two levels, the first is image class annotation where the images are sorted into the 12 classes based only on the object instances regardless if the object is the dominant majority in the image. Second is bounding box annotation of the objects, where every instance of any of the 12 classes are annotated in all images using Piotr's Computer Vision Matlab toolbox[1] (Dollár (n.d.)). Piotr's toolbox provide an easy to operate user interface for multi-class object labeling as shown in

[1]https://pdollar.github.io/toolbox/

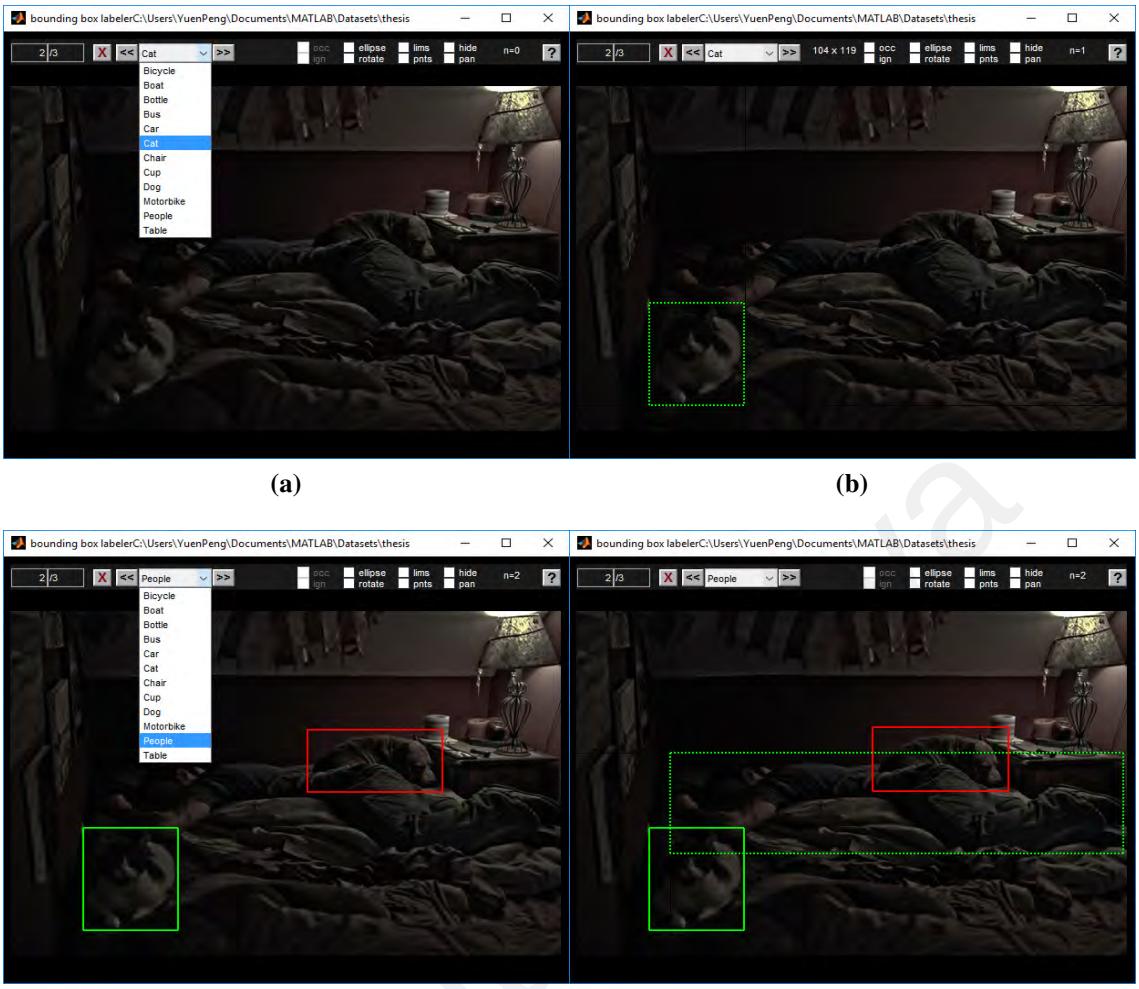


Figure 3.4: Object annotation using Piotr's Toolbox. To annotate, the object class is first selected (a), then the bounding box is drawn by clicking and dragging from the top left corner then bounding the object (b). The process is repeated for multiple objects in the same image as shown in (c) and (d).

Fig. 3.4. The toolbox stores the upper left coordinate (x, y), width (w), and height (h) in pixels, of the drawn bounding box and the corresponding object class into a *txt* file that allows easy read and write (See Appendix B for examples).

Figures 3.5 - 3.7 show the statistics of the image amount and fraction with respect to the annotations. Most of the images provide a single instance of the object as shown in Figure 3.6a, but a considerable amount of the images has more instances with the maximum number of bounding box annotation found in an image is 58, as shown in Figure 3.6b. Images that contain multiple instances can be a mixture of different objects, as shown in Figure 3.6b as well. While the number of images in the image level annotation

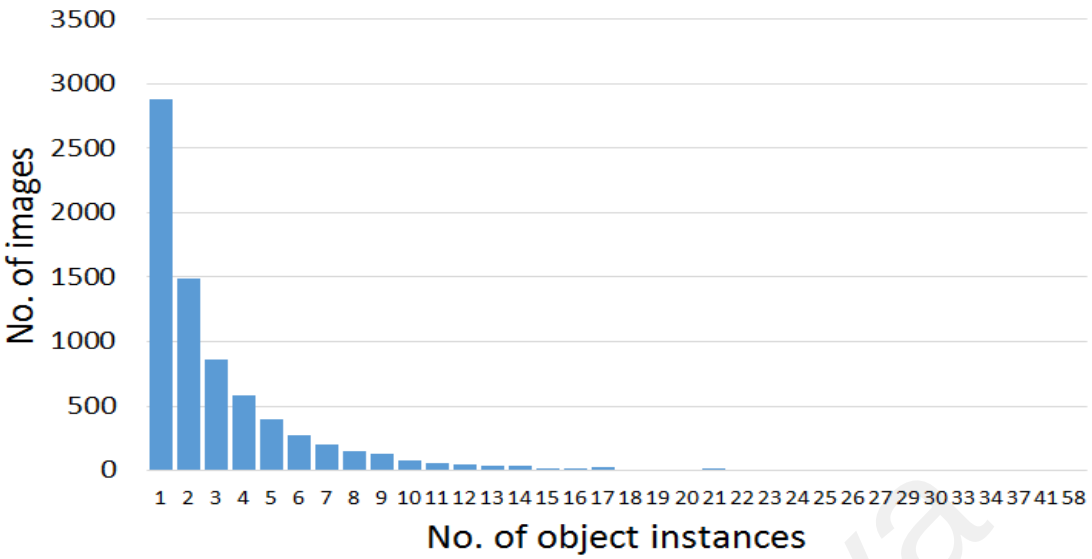
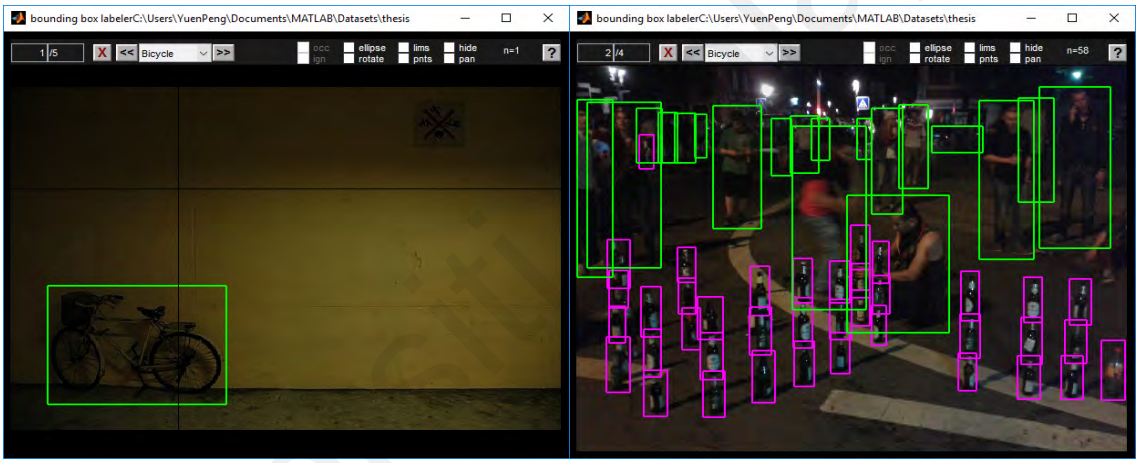


Figure 3.5: Object instances per image of ExDark data.



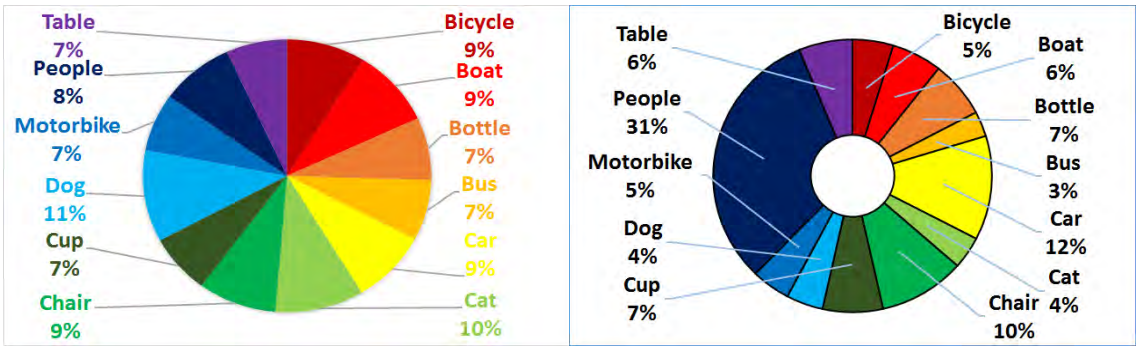
(a) **(b)**

Figure 3.6: Example (a) image with least amount of object annotated, and (b) image with the most objects annotated.

are kept relatively balanced as shown in Figure 3.7a, most of the bounding box annotations are from the *People* class, as seen in Figure 3.7b and the examples shown in Fig. 3.8. In the total of 23,710 object instances annotated, there are 7,460 *People*, from single person to a crowd, thus, this would be useful for pedestrian detection work as well.

3.2.3 Types of Low-light

From the collection of data, it was also identified that there are 10 types of low-light conditions, of both indoor and outdoor environments, commonly captured in images.



(a) (b)

Figure 3.7: (a) Fraction of image classes and (b) Object occurrence in ExDark dataset.

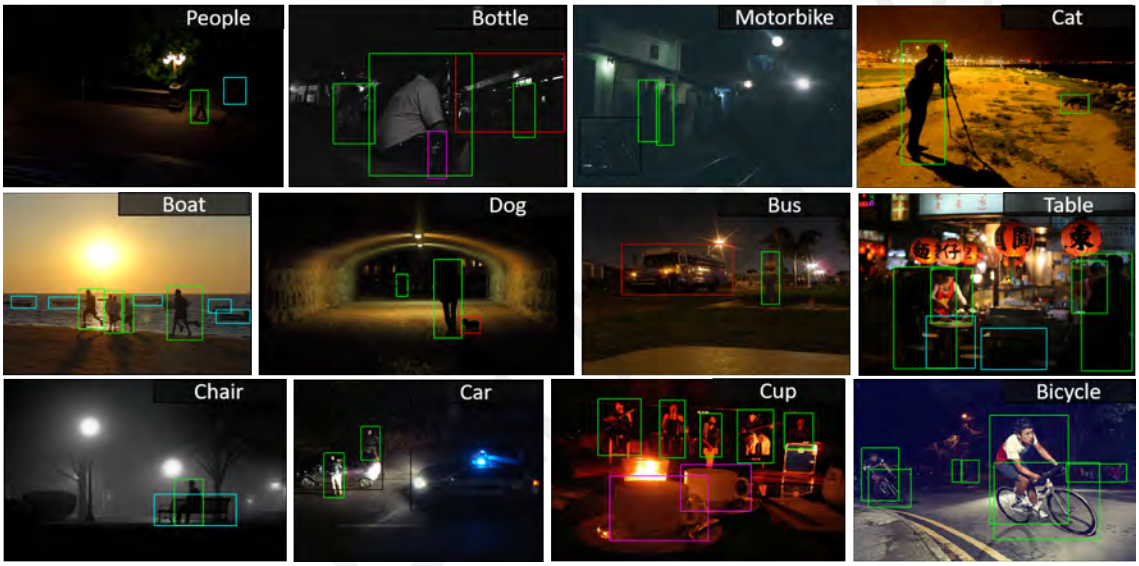


Figure 3.8: Examples of images from every class containing the *People* object.

These image types are established based on observation of common characteristics found in groups. Examples of the types are shown in Figure 3.9 and explained as follows:

- **Low**: Images with very low illumination and hardly visible details.
- **Ambient**: Images with weak illumination and the light source is not captured within.
- **Object**: Images where there is/are brightly illuminated object[2](s) but surroundings are dark and the light source is not captured within.
- **Single**: Images where a single light source is visible.

[2]The illuminated object is not necessarily from the 12 specified classes.

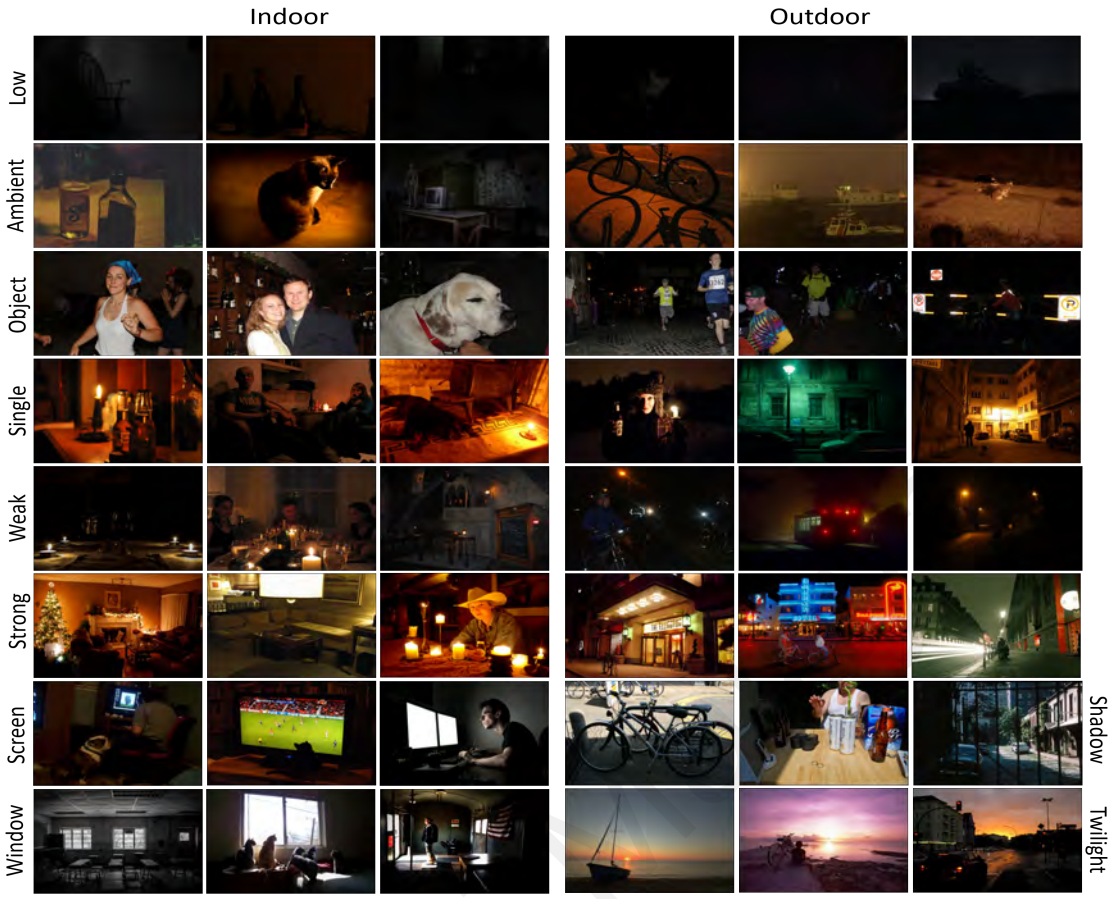


Figure 3.9: Examples of low-light image types in ExDark.

- **Weak**: Images with multiple visible but weak light sources.
- **Strong**: Images with multiple visible and relatively bright light sources.
- **Screen**: *Indoor* images with visible bright screens (i.e. computer monitors, televisions).
- **Window**: *Indoor* images with bright windows as light sources.
- **Shadow**: *Outdoor* images captured in daylight but the objects are shrouded in shadows.
- **Twilight**: *Outdoor* images captured in twilight (i.e. time of day between dawn and sunrise, or between dusk and sunset).

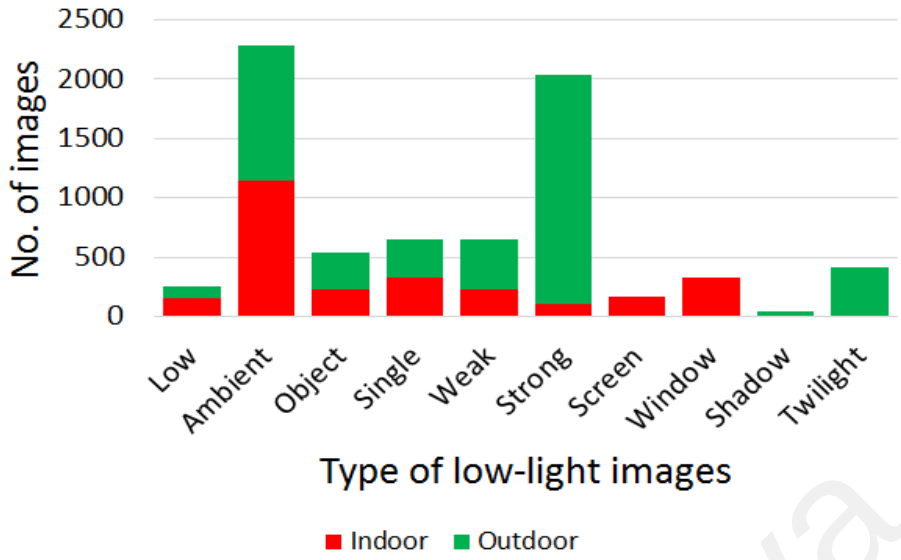


Figure 3.10: Statistic of image illumination types found in ExDark.

This categorization of low-light images has not been done before in any datasets hence it would be valuable for future research, particularly low-light image enhancement, as identifying the different illumination types could assist in the design of enhancement algorithms to handle the over and under enhancement problem accordingly. Figure 3.10 shows the statistics of the different illumination types found in the dataset, and further examples of the images from the ExDark can be found in Appendix C

3.3 Summary

This chapter has detailed the progression of object image datasets from earlier times of computer vision research till now which prompt the next evolution of the object dataset. This is followed by the information regarding the proposed ExDark low-light object image dataset, including data collection procedures and sources, annotation tool and process, as well as statistics of the data annotated.

The proposed ExDark contains 7,363 low-light images, encompassing 10 types of low-light conditions. 12 object classes were annotated for each of the images, from image level having one class represent the whole image, till object level using bounding box to indicate the location of the object in the image. This is the first dataset of its kind, thus

can act as a pioneering benchmark for both low-light image enhancement and object detection research works. The dataset is available to the public at *https://github.com/cs-chan/Exclusively-Dark-Image-Dataset*.

CHAPTER 4: ANALYSIS OF LOW-LIGHT IMAGES

From investigations of notable datasets, it is found that low-light is commonly glossed over in object dataset analyses (Everingham et al. (2015); Lin et al. (2014); Russakovsky et al. (2015)) with the preferred emphasis on object instances, scale, occlusion, and quantity. In consequence, the state-of-the-art object detectors, past and present (Felzenszwalb et al. (2008); He et al. (2016); Krizhevsky et al. (2012); Simonyan & Zisserman (2014); J. Wang et al. (2010)), are neither designed nor were they analyzed on low-light, given the samples they had to work with. This has also indirectly led many to oversimplify the diversity and challenges of low-light. Considering very early computer vision works, such as well-known feature extractors (Dalal & Triggs (2005); Lowe (2004)), had already strove for illumination invariance in their designs, it is understandable that many would consider illumination or low-light as just an auxiliary element to other challenges without going into a deeper understanding. Particularly, with the emergence of deep learning, machine learning is expected to be able to counteract this problem with ease.

A crucial belief in this work is that the characterization of low-light as just "illumination variation" does not fully define the challenges as the "variations" encompass much more. For example, low-light condition can emerge depending on the time of day (e.g. twilight, nighttime), location (e.g. indoor, outdoor), and the availability of light sources and their types (e.g. the sun, man-made lights). The combination of these three factors can create a great deal of disparity between image to image or even within an image itself. The impact of these variations has been left unexplained in most works, especially in object detection, however a grasp of their behavior can potentially advance the field. Though, rather than disregarding the milestones of researches so far, the belief is simply that a gap has been overlooked in the common object data and analysis. Thus, this chapter details the analytical efforts of this work in gaining a better understanding of the low-light

phenomenon in regards to computer vision.

4.1 Data

Before carrying out any analysis, the data which is to be studied needs to be established. For this work, the proposed ExDark image dataset is used, and in order to establish a comparison, the MSCOCO is chosen as the baseline dataset for its challenging nature to represent the current trend of object datasets that are dominated by bright images. However, since the ExDark is considerably less in amount, images were sub-sampled from the MSCOCO for the study. The sub-sampling was done on the training and validation sets of the MSCOCO only as the annotation for these sets were provided. The criteria for an image to be extracted is that it contains at least one of the 12 object classes of the ExDark dataset irregardless of the presence of other object classes.

Generally, 150 images per class were extracted from the validation set and 500 images from the training set, except for the classes where there are insufficient images containing the specified object, hence all the images were taken for such cases. Additionally, following the nature of the images collected for the ExDark where similar objects are categorized into a coarse class, the *Bench*, *Chair* and *Couch* classes from MSCOCO were merged into a single *Chair* class in the subset, likewise for the *Wine glass* and *Cup* classes, merged as only *Cup*. As for the annotations used for the analysis work, only the bounding boxes of the 12 object classes were taken, and the image level annotation is set based on the highest object instances found within a specific image. Table 4.1 shows the number of images per class of the ExDark and the sub-sampled MSCOCO. There are a total of 23,710 object instances found in the ExDark, whereas the MSCOCO subset has 34,370 object instances.

Table 4.1: Number of images per object class used for analyses.

Dataset	ExDark	MSCOCO
Class	Number of Image	Number of image
Bicycle	652	603
Boat	679	650
Bottle	547	650
Bus	527	564
Car	638	650
Cat	735	650
Chair	648	651
Cup	519	650
Dog	801	650
Motorbike	503	644
People	609	650
Table	505	650
Total	7,363	7,662

4.2 Analyses

The objective of analysis is to gain a better understanding of the subject matter and subsequently develop an optimal solution based on the findings. Therefore, it is essential for the medium of analysis to bring forward the appropriate aspects to achieve the specified goals. In this study, the aim is to understand the effects of the low-light phenomenon in images on applications, particularly object detection. Hence, the analysis is distinguished into two approaches, low level and high level analyses.

Low-level analysis looks into the causes of the low-light phenomenon as well as the characteristics of pixel intensities of the captured images by studying the global intensity distributions and also local region intensity variations. On the other hand, high level analysis analyzes the performance of features commonly used in object detection works when applied on low-light conditions. Specifically, object proposal algorithms that make use of hand-crafted features (Cheng et al. (2014); Fang et al. (2016); Zitnick & Dollár (2014)), and object classification CNN (He et al. (2016)) that learns features, were employed on both low-light and bright images for a comparative study.

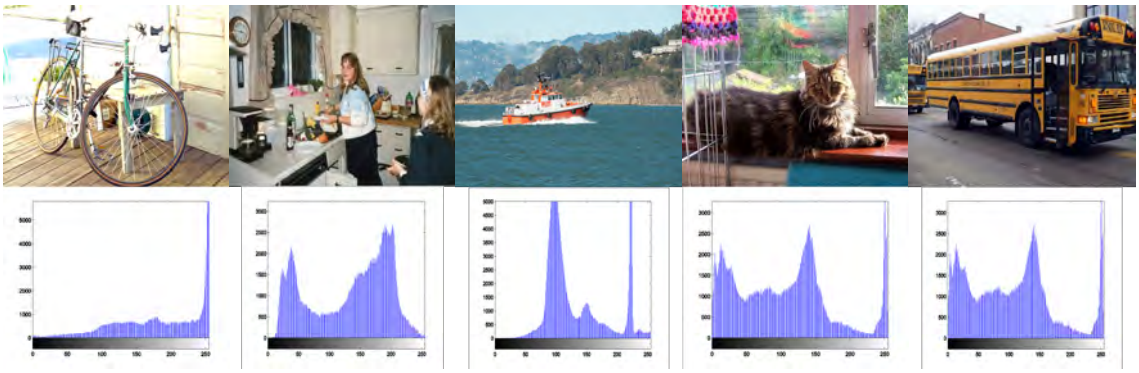


Figure 4.1: Bright images and their global intensity histograms.

4.2.1 Low Level Analysis

The scene luminance of images captured in bright environments are typically very high due to bright sources of light, such as the sun, that is able to encompass large areas. As shown in Fig. 4.1, the content and details of the images are clearly perceivable due to ample lighting, and the respective intensity histograms are well distributed across the intensities. On the contrary, low-light images have low illumination where the appearance of the captured object(s) lacks details and may look invisible as shown in Fig. 4.2. Moreover, their intensity histograms are greatly biased towards the lower levels. This is because the lighting in low-light environments are provided by limited sources that are comparatively weaker, such as the setting sun (twilight), street lights, or car lights.

Looking into the causes of the many variations, it is found that the illumination differences in low-light environments are highly dependent on the light sources where they affect the luminance level of an image both globally and locally. For example, images that are taken at dusk, dawn, and nighttime, either outdoor or indoor, each has different levels of light from one another. Furthermore, artificial light sources are used in such environments which adds more variations, like street lights, table lamps, fluorescent lamps, and LED bulbs to name a few. This diversity of light sources dictates the intensity that reaches an object, subsequently affecting the image's overall intensity and contrast captured by a camera. This is reflected in the intensity histograms obtained from both the MSCOCO

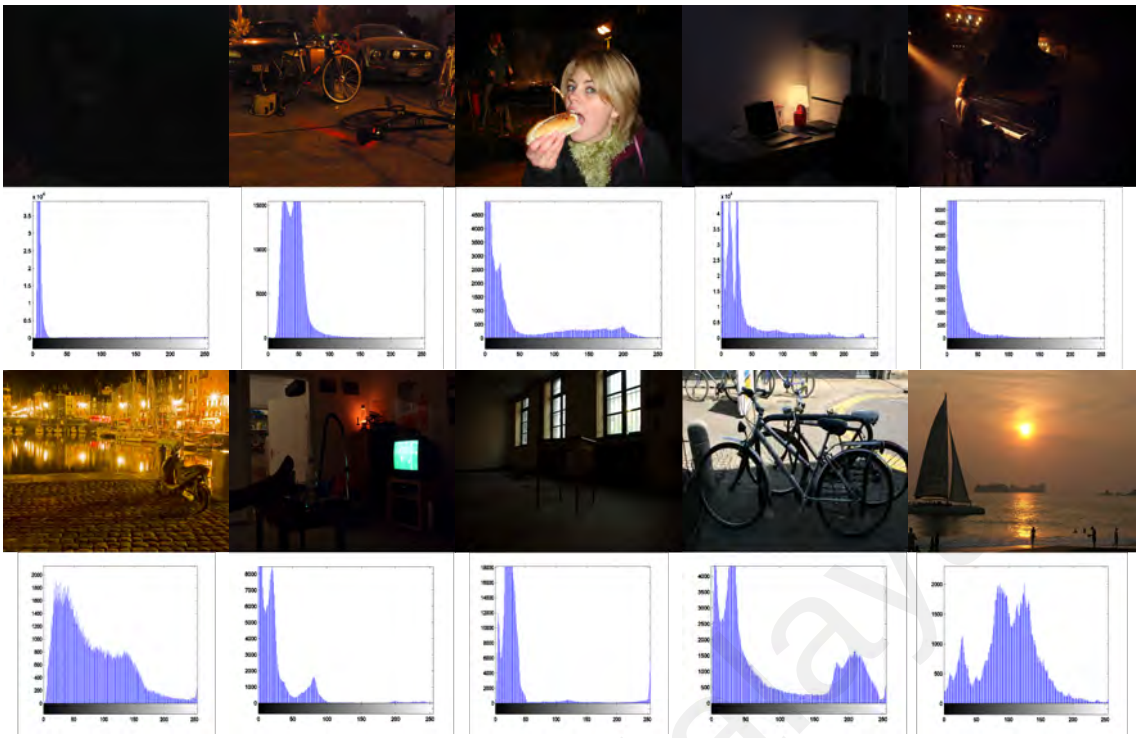


Figure 4.2: Low-light images and their global intensity histograms. Low light image types, from upper left: Low, Ambient, Object, Single, Weak, Strong, Screen, Window, Shadow, and Twilight.

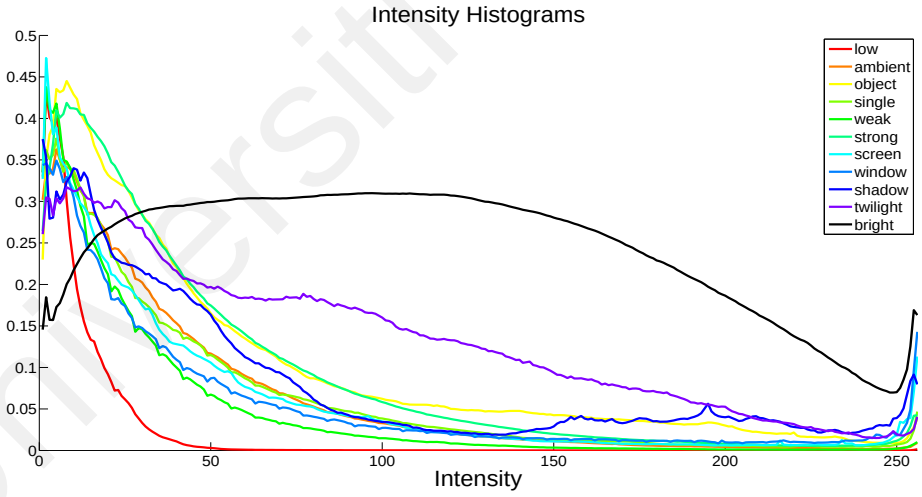


Figure 4.3: Average global intensity histograms of the MSCOCO subset (bright), and the ExDark averaged based on the lighting types.

subset and ExDark shown in Fig. 4.3 where different lighting types has subtle differences between one another, while bright images histogram greatly deviates from them. Therefore, this image to image difference is termed as **global illumination variation**.

On the other hand, in very low-light environments such as nighttime, objects are only apparent when near to the light source but become increasingly obscure as they

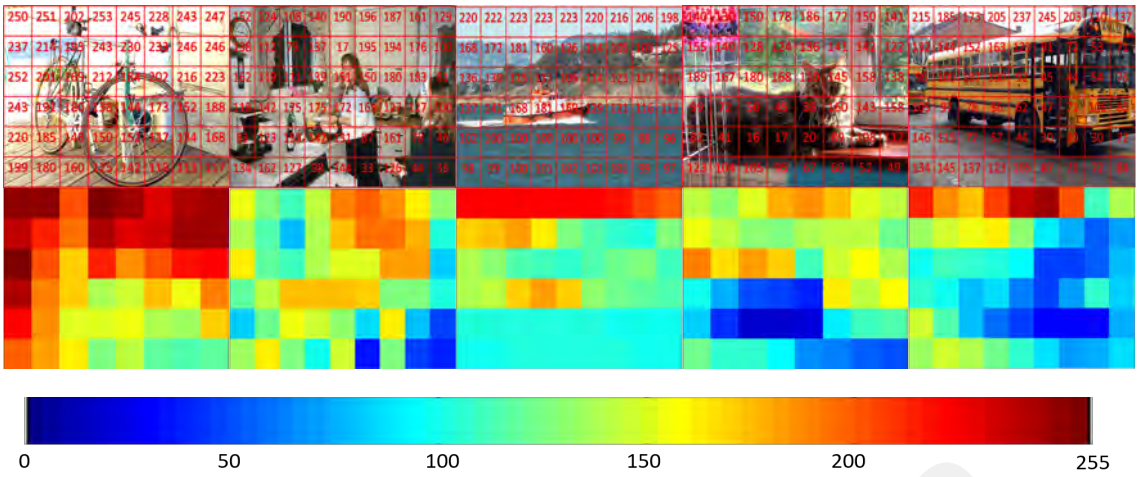


Figure 4.4: Average intensity values of local patches and heat maps of bright images from Fig. 4.1.

move further away. This brings considerable illumination variations within one image. Figures 4.4 and 4.5 show the patch intensity values and illustrated as heat maps of bright and low-light images corresponding to Fig. 4.1 and 4.2 respectively. In Fig. 4.4, the intensity values are relatively high and consistent throughout the image. However, in Fig. 4.5, it can be seen that the average intensities are in the much lower range and reduce as the patches move away from the light source with very large differences, as illustrated by the heat maps. Furthermore, when there are more than one light source available, the intensity variation becomes even more severe, as seen in the Weak, Strong, and Screen types in Fig. 4.5. Hence, these image specific variations are termed as **local illumination variation**.

Based on this analysis, it is noted that both the global and local illumination variations requires specific attention in order to progress the development of a practical low-light system.

4.2.2 High Level Analysis

Many higher level computer vision works, such as feature extractors and object detectors, claim robustness in the designs that would handle the illumination problem, though there

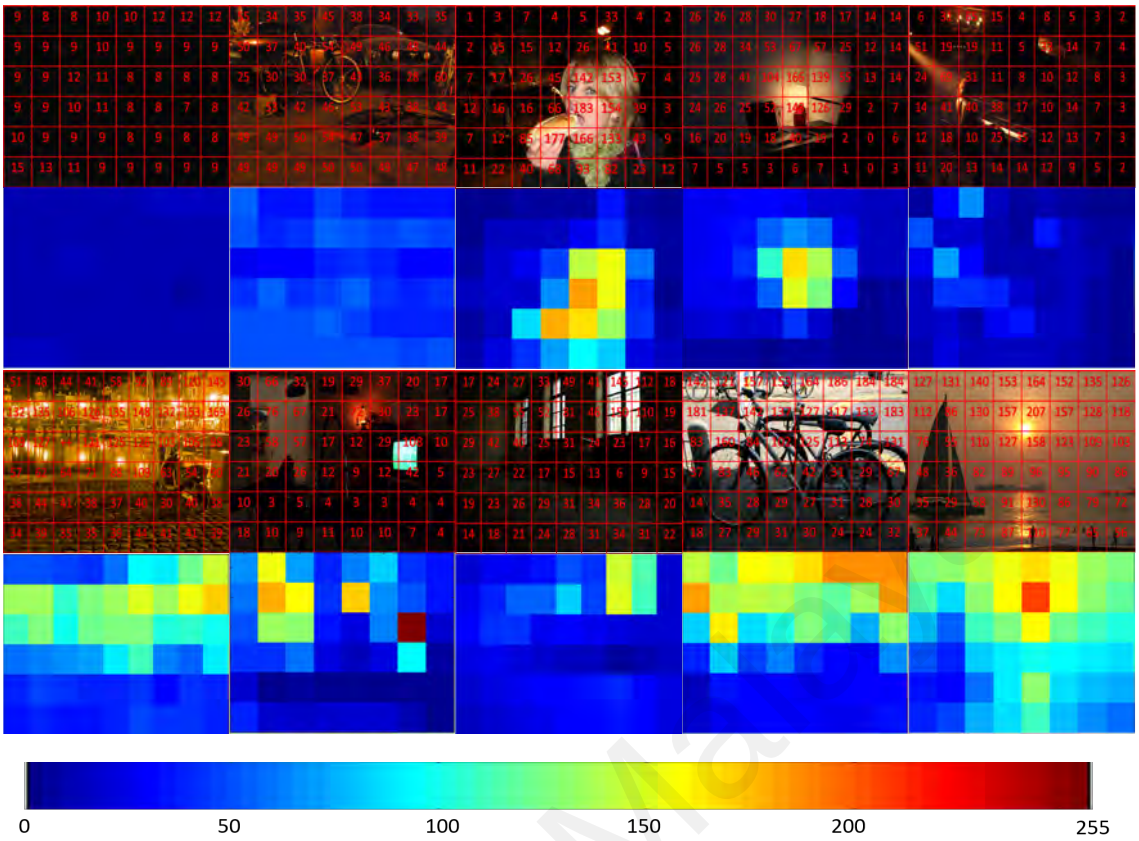


Figure 4.5: Average intensity values of local patches and heat maps of low-light images from Fig. 4.2. Low light image types, from upper left: Low, Ambient, Object, Single, Weak, Strong, Screen, Window, Shadow, and Twilight.

were no explicit assertion about their capability addressing the global and local illumination variation of low-light environments. Hence, the high level analysis in this section serves to look into the effectiveness of existing object features, both hand-crafted and learned, in addressing the low-light challenges.

4.2.2 (a) Performance of Hand-Crafted Features

Hand-crafted features are designed computations to extract meaningful information, based on established insights on the behaviors of the image contents, as opposed to learned features where computational models are trained to discover the meaningful information by themselves. While the progress of deep learning in these few years has seen a shift in preference towards learned features, hand-crafted features are still employed, particularly for the object proposal task due to their high speed and low complexity nature. The inten-

tion of this analysis is to look into the abilities of classical hand-designed features when handling object detection in low-light images, thus algorithms that use different types of features are engaged for the comparison, namely Edge Boxes[1] (Zitnick & Dollár (2014)), Binarized Normed Gradients (BING)[2] (Cheng et al. (2014)), and Adobe Boxes[3] (Fang et al. (2016)), instead of deep learning based proposers (Redmon et al. (2016); Ren et al. (2015)). Descriptions of these methods are as follows:

- **Edge Boxes**, as stated in the name proposes object bounding boxes by grouping *edges*, and uses the edge inside and overlapping the bounding box to compute a score indicating the likelihood of an object, i.e. objectness. Given a bounding box B_i in an image, $w_B(s_j) \in [0,1]$ measures if an edge s_j is fully inside or overlapping B_i based on the affinity between edges. $w_{B_i}(s_j) = 1$ indicates that the edge is fully inside the box whereas edges overlapping or outside B_i is shown by $w_{B_i}(s_j) = 0$. The objectness score is computed by

$$h_{B_i} = \frac{\sum_j w_{B_i}(s_j) m_j}{2(B_i^w + B_i^h)\kappa}, \tag{4.1}$$

  where m_j is the sum of edge pixel magnitudes, B_i^w and B_i^h are the width and height of B_i respectively, and κ is the box size offset. A higher h_{B_i} denotes a higher possibility that B_i encloses an object. Bounding boxes are searched by sliding windows where the result is a set of bounding box proposals $B = \{B_i | i = 1, \ldots, n\}$.

- **BING** is based on the correlation between object boundaries and norm of image *gradients*. The method uses a 64 dimension BING feature based on the Euclidean norm of the gradients in 8×8 regions of an image in multiple scales to evaluate ob-

[1] https://github.com/pdollar/edges

[2] https://github.com/tfzhou/BINGObjectness

[3] https://github.com/fzw310/AdobeBoxes-v1.0-/tree/master/AdobeBoxes(v1.0)

jectness. Initially, object boxes are searched by scanning over the image at different scales using windows with predefined sizes and aspect ratio. Each of the windows are scored using a linear SVM model given the BING features. This SVM model is then trained using BING features of groundtruth object windows as positive samples and features from random background as negative samples. The classified object windows are then refined using Non-Maximal Suppression (NMS)) to remove windows less likely to contain objects, for example window size of 10×500 is less likely an object as compared to 100×100 sized windows. The remaining windows are then scored again using another linear SVM model that takes the score from the previous model as input. This score is called the calibrated filter score, where it calibrates the score from the previous model to get the final objectness score.

- **Adobe Boxes** uses groups of *superpixels* with high contrast from the background as the representation of object parts, named adobes, to propose object bounding boxes. The spatial concentration of adobes are used to calculate the objectness score, that is, the closer or more compact the adobes, the more likely they constitute an object. To extract object adobes, an image segmentation method is used to obtain superpixels and histogram intersection distance is used to measure the distance between the superpixels. Given an initial window, the following sets of superpixels are determined:
  - Background superpixel set, S_b where the superpixels intersect with the window.
  - Internel superpixel set, S_i where the superpixels are located inside the window and do not touch the boundary.
  - Seed superpixel set, S_s where the superpixels are likely to belong to the object. This is determined by calculating the local contrast of S_i and selecting the ones

with highest probability.

- Candidate adobe set, S_c where the superpixels are inside of the window and may touch the boundary, hence $S_i \sqsubseteq S_c$.

The final object adobe set of superpixels, S_o is then determined from the S_c by comparing its contrast to the superpixels in S_b and S_s sets. If the contrast distance of a superpixel $s \in S_c$ is closer to S_s, the superpixel is an object adobe. Once the object adobes are found, the window is refined to obtain the object bounding box. The bounding boxes are then ranked using the adobe compactness which considers that the more compactly the object adobes are spatially distributed, the more likely it is for the box to capture the object. The score is defined as

$$AC(B) = \frac{\sum_{s \in S_o} |s|}{|B|} \tag{4.2}$$

where s is a superpixel and B is the bounding box. For Adobe Boxes with the same adobe coompactness, the larger bounding box is favored because it indicates that the object is more salient and more parts were captured. This method can also be used to refine proposals produced by other methods, e.g. the paper showed that it works well when combined with BING (AdobeBING).

Quantitative Evaluation. This evaluation is to assess the ability of hand-crafted features to detect objects in both bright and low-light images, disregarding the identity of the objects. Experiments were performed to compare the detection rate (detections/groundtruths) and recall (detections/proposals) between the datasets using each proposal method. In the tests, the methods were set to produce a maximum of 1000 bounding boxes, however the amount could be less depending on the algorithms ability to confidently propose the boxes. As for the evaluation, the Intersection over Union (IoU) metric

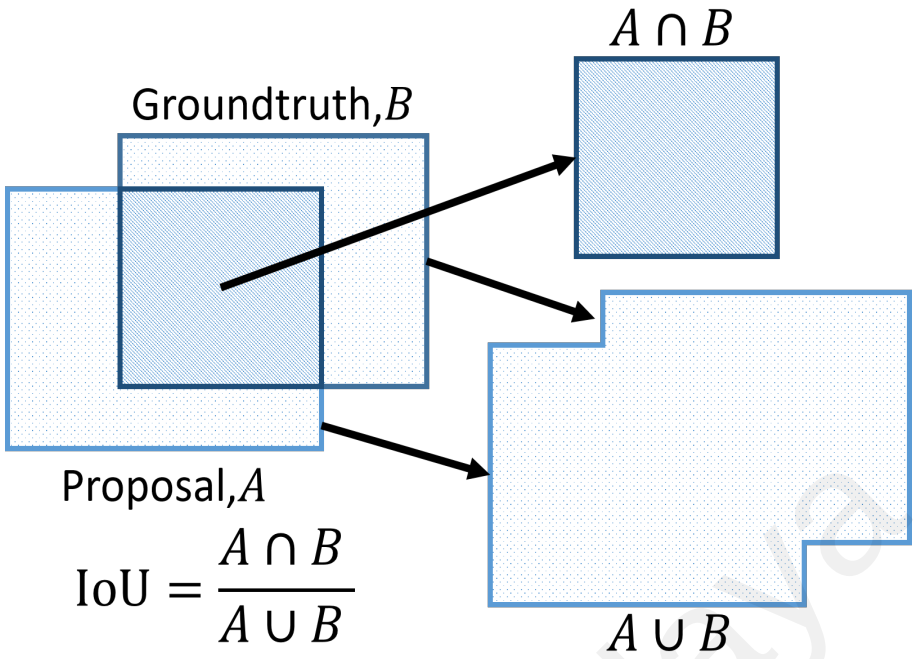


Figure 4.6: Illustration of IoU computation.

is used, where it calculates the overlap of the bounding boxes by dividing the area of intersection between the proposals with the groundtruth boxes with the area of their union, as shown in Fig. 4.6. In this analysis varying thresholds, from 0.5 to 1.0, were tested.

Implicitly, as the IoU increase, the detection rate and recall will reduce as the criteria to constitute a detection becomes stricter, as seen in performance graphs in Fig. 4.7. At lower IoU, the detection rate is higher for images from the ExDark but this condition gradually inverts as the IoU increases. From the onset, the higher detection rate on the ExDark seems to indicate more object detections, however, the results in Table 4.2 show that the average detection in low-light images are less than MSCOCO for all methods. Hence, it is postulated that the reason for the higher detection rate is caused by the number of groundtruth where the images in MSCOCO contain more objects that remain undetected. These undetected objects can be attributed to the complexity of the MSCOCO images where many of the objects are too small, occluded, or only partially shown in the image, a common trait in challenging bright datasets. Whereas the images from ExDark mostly contain the full objects where the main challenge comes from the illumination. Nonethe-

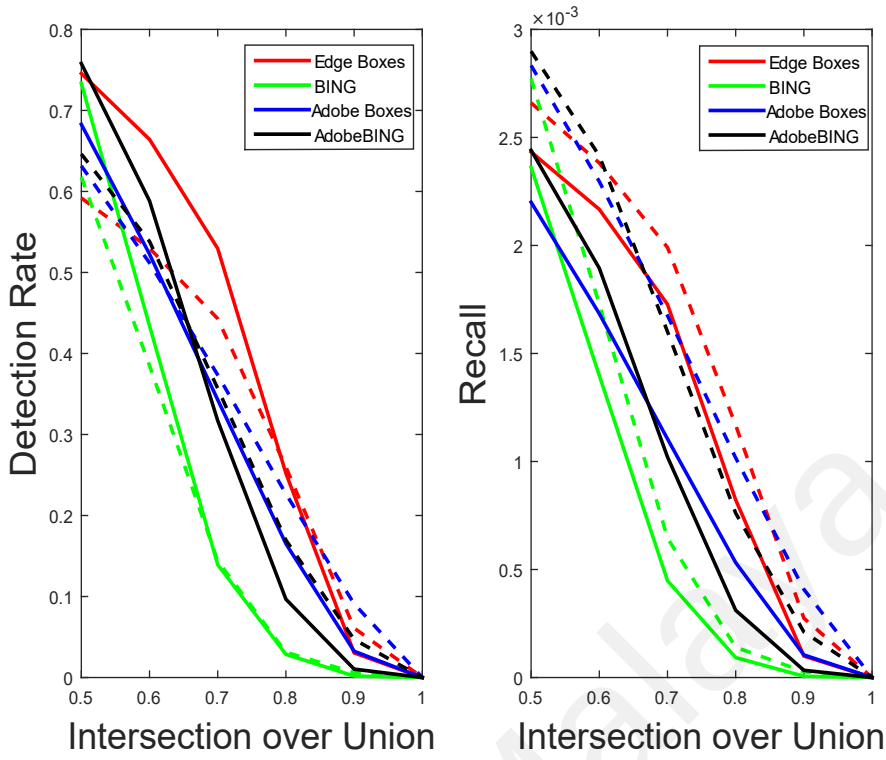


Figure 4.7: Detection rate and recall of Edge boxes, BING, Adobe boxes, and BING refined by Adobe boxes (AdobeBING), at maximum proposal of 1000 boxes. (The solid lines are the performance on ExDark, and the dotted lines shows the performance on MSCOCO)

Table 4.2: Average proposals, average detections, detection rate, and recall of tested proposal methods at maximum proposal of 1000 and IoU of 0.7.

Methods	Dataset	Avg. Proposal/im	Avg. Detection/im	Detection Rate	Recall
Edge Boxes	MSCOCO	998	**1.9871**	0.4430	**0.0020**
	ExDark	987	1.7050	**0.5295**	0.0017
BING	MSCOCO	1000	0.6457	0.1439	0.0006
	ExDark	1000	0.4483	0.1392	0.0004
Adobe Boxes	MSCOCO	1000	1.6753	0.3735	0.0017
	ExDark	999	1.1039	0.3428	0.0011
Adobe BING	MSCOCO	1000	1.6010	0.3569	0.0016
	ExDark	1000	1.0209	0.3170	0.0010

less, the low detection rate for ExDark at higher IoU is also an indication that it is more challenging to get an accurate localization in low-light images as compared to the bright images.

On the other hand, the recall on ExDark is obviously lower than the MSCOCO data using any of the methods. This result infers that most of the proposals in the low-light images are not valuable, even though the average proposal per image maybe lower than that in MSCOCO, such as for the Edge Boxes and Adobe Boxes in Table 4.2.

Qualitative Evaluation. Further study of the results of different features were conducted by investigating some of the qualitative examples of both bright and low-light images in Fig. 4.8 and Fig. 4.9 respectively, as well as visualizations of the features used by the proposers.

In Fig. 4.8, it can be seen that the MSCOCO images has objects that are very small compared to the image size, which causes the methods, particularly the Edge Boxes and BING to fail. It can be understood by examining their respective edge and gradient images, that the features are unable to capture the details of really small objects. On the other hand, Adobe Boxes and AdobeBING are better as superpixels are more precise in segmenting the objects from background, but it still could not fully solve the problem as seen in the undetected small objects in the last example of Fig. 4.8.

On the contrary, the failures in the ExDark are not due to the object scale, but from factors related to low-light, as shown in Fig. 4.9. The first is the additional noise in low-light images that causes the failure due to interference from extra features, as seen in the first two rows of images in Fig. 4.9. Even for successful proposals, the alignment is rather far from the groundtruth. These noises are usually caused by the high camera ISO setting used to compensate the low-light level but at the same time it makes the camera oversensitive to the surrounding light. The other cause is the blending of the objects either to the background or to other objects, as seen in the last two rows of examples in Fig. 4.9. The methods are especially weak for these types of conditions because the gradient boundaries are unclear and the superpixels were unable to distinguish the difference between the low valued pixels of objects and backgrounds.

Further Look into Low-light. The detection and recall of the methods were separated into the 10 types of low-light images that have been established for a further investigation into the effects of the results based on illumination. Figures 4.10 and 4.11 show the detection rate and recall respectively, where Edge Boxes performs the best for

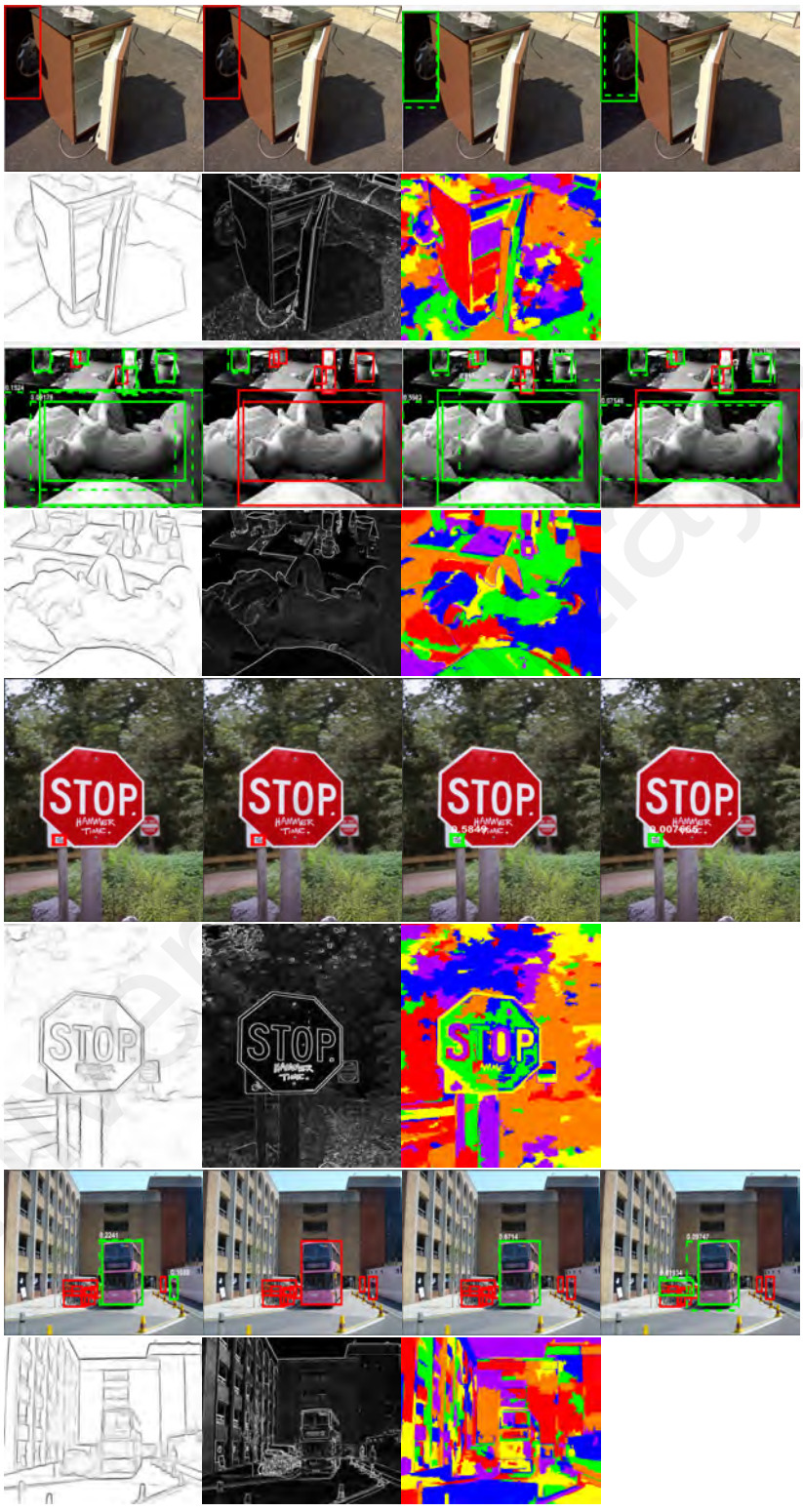


Figure 4.8: Examples of proposals on MSCOCO images and visualizations of their respective features. (Red: undetected groundtruth; Green: detected groundtruth, Green dotted: proposed box) From left: Edge Boxes, BING, Adobe Boxes, and AdobeBING. (Maximum proposals = 1000; IoU = 0.7)

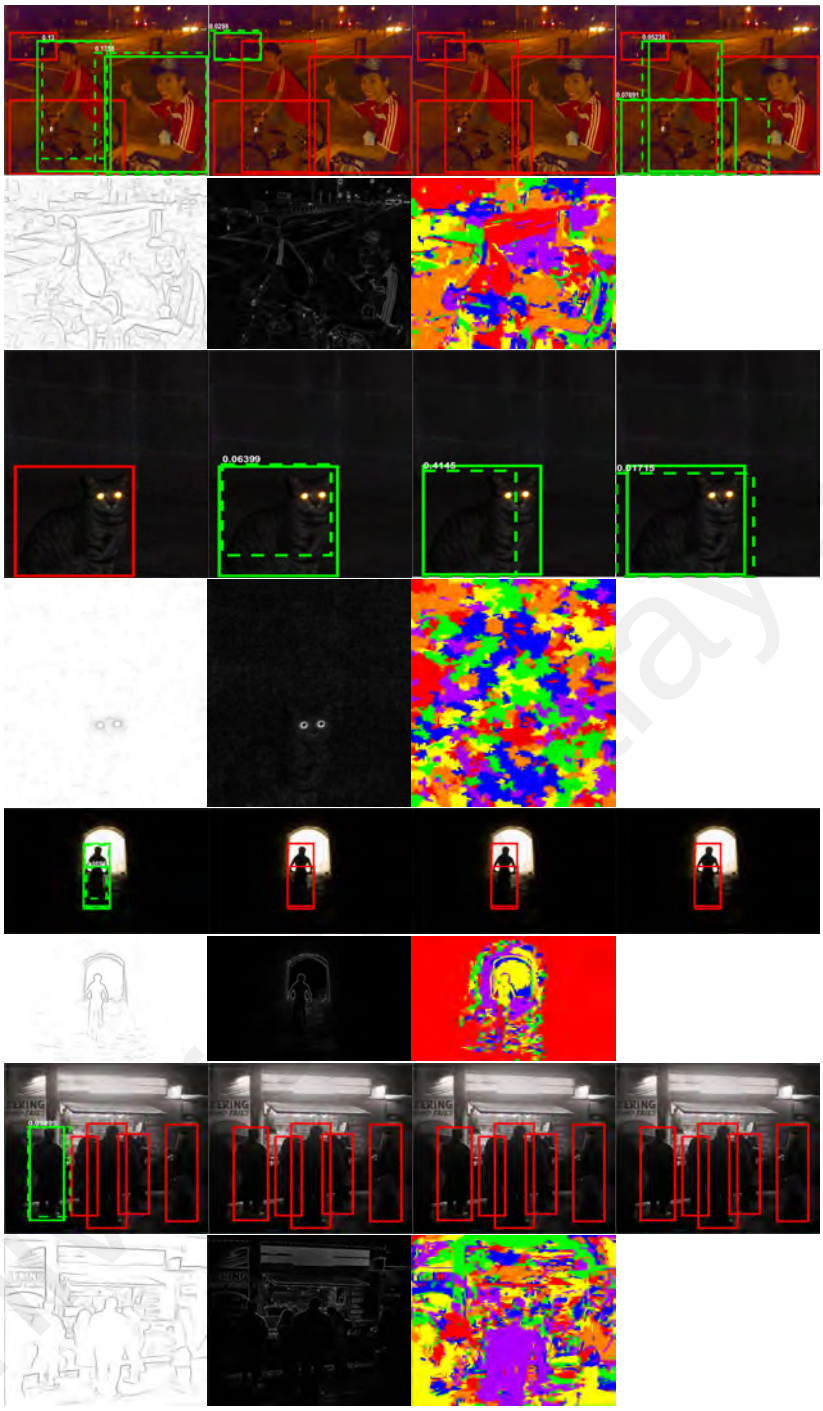


Figure 4.9: Examples of proposals on ExDark images and visualizations of their respective features. (Red: undetected groundtruth; Green: detected groundtruth, Green dotted: proposed box) From left: Edge Boxes, BING, Adobe Boxes, and AdobeBING. (Max. proposals = 1000; IoU = 0.7)

all types of low-light conditions. Images with Ambient and Single lighting show the best detection rate, while Low and surprisingly Strong lighting are the weakest. Whereas for the recall, the Object lighting type is the best while Low is the least. Figure 4.12 shows examples of Edge Boxes detections in the different types of lighting.

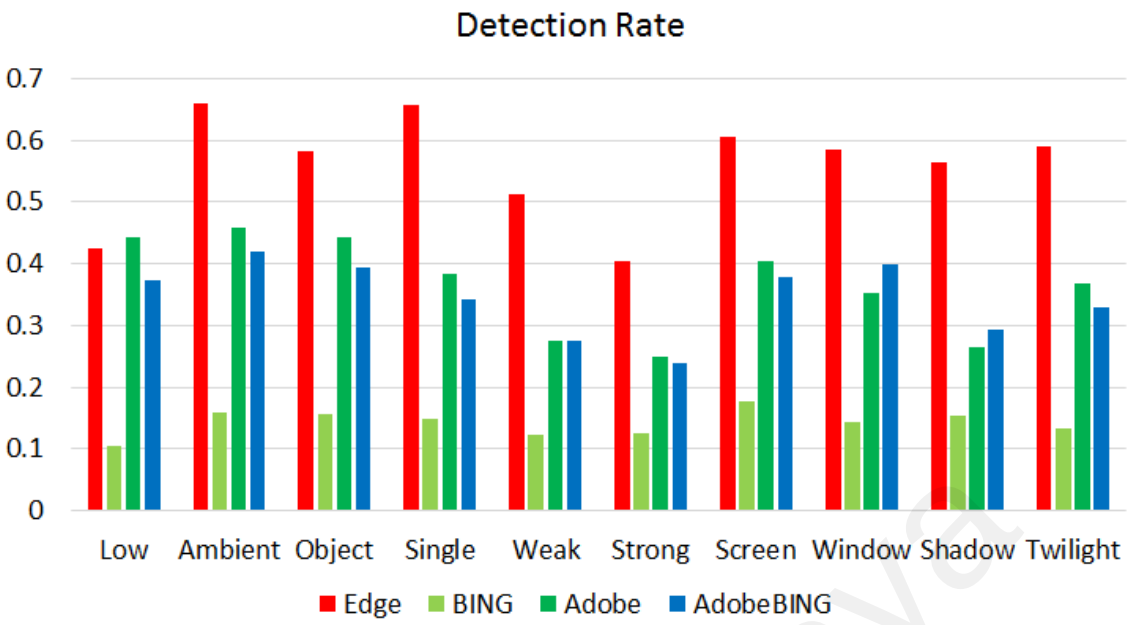


Figure 4.10: Detection rate of Edge boxes, BING, Adobe Boxes, and BING refined by Adobe boxes (AdobeBING), sorted into low-light image types. (Maximum proposals = 1000; IoU = 0.7)

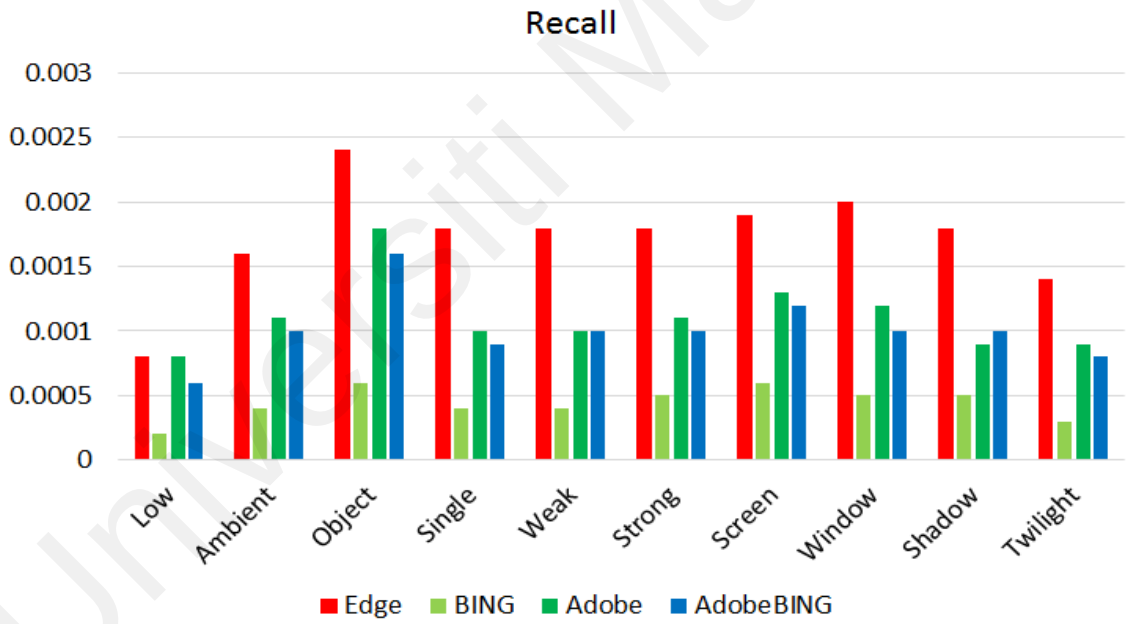


Figure 4.11: Recall of Edge boxes, BING, Adobe Boxes, and BING refined by Adobe boxes (AdobeBING), sorted into low-light image types. (Maximum proposals = 1000; IoU = 0.7)

The method performs quite well for the Ambient and Single light types because there are still enough light in the image to highlight the object features, particularly when the objects are nearer to the source of light. Whereas for very low light images, the objects are more likely to blend into the background. On the other hand, images taken in strongly lit low-light environments are expected to show more features, however, such environments

Figure 4.12: Examples of Edge Boxes proposals (Max. proposals = 1000; IoU = 0.7) on different types of low-light images and visualizations of the edge features. (Red: undetected groundtruth; Green: detected groundtruth, Green dotted: proposed box) From left, first row: Low, Ambient, Object, Single, Weak; second row: Strong, Screen, Window, Shadow, Twilight.

are also more cluttered with objects and irregular light sources that results in complex images, subsequently deteriorating detection performance.

Considering the recall, very low light images have the lowest value because the contrast of the objects are either too low for the object features to be extracted or the image is saturated with noise due to the camera's high ISO setting. Images with a well illuminated object but low-light surroundings give the best recall because the well lit object will mostly be detected even if the other objects in the low-light background are missed, hence aiding in the recall evaluation. For the most part, the detection rates using these hand-crafted approaches are below 70% for any type of low-light conditions, which leaves room for improvement before a good low-light object detection system can be achieved.

4.2.2 (b) Insights from Learned Features

After exploring the performance of hand-crafted features in Section 4.2.2 (a), here the capabilities of learned features in low-light are tested. In contrast to hand-crafted features, learned features rely on the computation of machine learning algorithms to uncover the best representations for a given task. At first, the features learned largely remain unknown as the high dimensional representations generated by machines could not be fully comprehended. Nevertheless, many works have since visualized high dimensional data and features (Donahue et al. (2014); Lee et al. (2017); Mahendran & Vedaldi (2015); Yosinski et al. (2015); Zeiler & Fergus (2014)) to understand and find out what the machines "see".

This section details the attempt to uncover the features in low-light images by visualizing a straight forward object classification CNN, as it has become the benchmark in learning discriminative features for the task. Specifically, the pre-trained Resnet-50 model (He et al. (2016)) was fine-tuned, on the MSCOCO and ExDark data, and evaluated their performance based on different ratios of bright and dark data used in the fine-tuning. Then, the behavior of the learned representations are studied in two ways. First off, the t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten & Hinton (2008); Van Der Maaten (2014)) is used to visualize a 2D mapping on the clustering behavior achieved by the learned feature vectors. The other is the visualization of the activations of convolution maps corresponding to the spatial location on the images (Yosinski et al. (2015)) in order to find out which part of an image "triggers" the classification outcome, i.e. the attention of the network.

Classification Performance. The Resnet model is chosen for this task because it is currently one of the top performing architecture in both the ILSVRC and MSCOCO challenges. In a plain CNN, convolution and pooling operations, i.e. layers, are sequentially stacked to form a network architecture that is trained with supervision to perform a task,

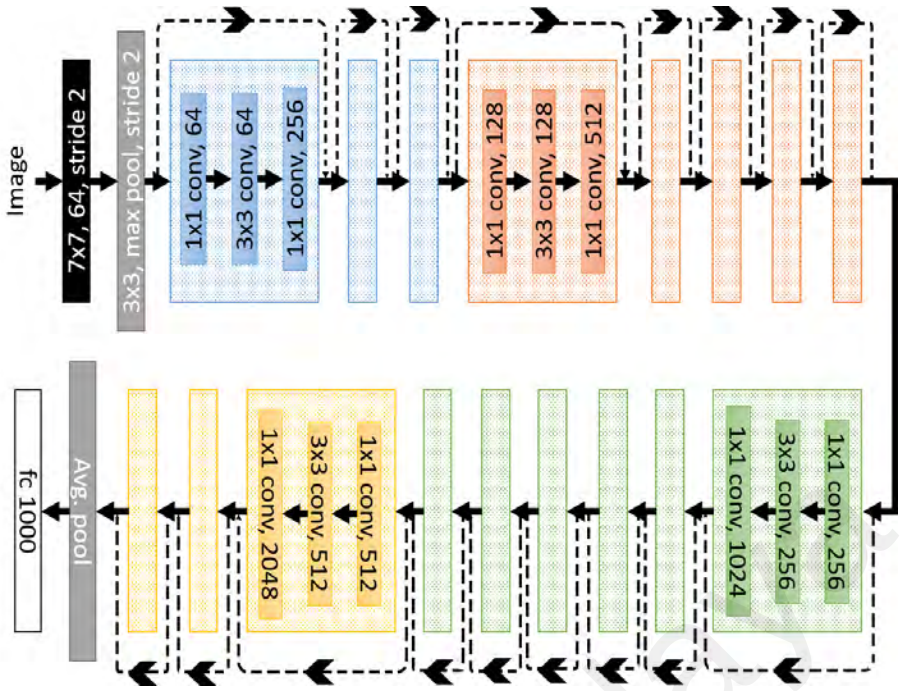


Figure 4.13: The residual network architecture with 50 layers (Resnet-50). The dotted lines show the shortcut connections that changes a regular CNN into a residual network. The shortcuts are implemented on every block containing 3 convolution layers, and there are 4 types of blocks (shown in varying colors) consisting different convolution parameters. Varying amounts of blocks are stacked with the shortcuts to form the network.

but it is able to learn features in an unsupervised manner. The Resnet is a CNN that implements residual function learning in the layers with respect to the inputs, i.e. a shortcut that bypass the layers as shown in Fig. 4.13. This architecture allows for easier network optimization of substantially deeper CNN models, from tens to hundreds of layers.

However, a common problem in training a CNN, especially one as deep as the Resnet architecture is overfitting, that is when the training data provided is not large enough the trained model is not generalized and performs poorly in testing (Donahue et al. (2014); LeCun et al. (2015)). Hence, on account that the amount of images in the ExDark is still too small to train a full CNN model from scratch, the task was approached by fine-tuning and the specific model chosen is the Resnet-50 that is pre-trained using ImageNet.

The Resnet-50 is chosen instead of its deeper counterparts to further minimize the chances of overfitting, whereas the ImageNet pre-training is able to initialize the network's weights to the same tasks, that is object classification. The fine-tuning scheme

Table 4.3: Accuracy of Resnet-50 models fine-tuned using different ratios of bright images (MSCOCO) and low-light images (ExDark). MSCOCO: performance on MSCOCO test images only, ExDark: performance on ExDark test images only, Overall: performance on test images of both sets.

	Training ratio	Test Accuracy		
Model	MSCOCO:ExDark	MSCOCO	ExDark	Overall
1	10:0	62.75%	43.15%	53.49%
2	9:1	**63.31%**	48.89%	56.50%
3	8:2	62.16%	52.75%	57.71%
4	7:3	61.25%	55.05%	58.32%
5	6:4	61.50%	55.64%	58.73%
6	5:5	61.18%	58.45%	**59.89%**
7	4:6	59.89%	58.99%	59.47%
8	3:7	58.00%	59.54%	58.73%
9	2:8	57.27%	61.45%	59.24%
10	1:9	55.38%	62.27%	58.64%
11	0:10	46.30%	**62.58%**	53.99%

introduces new classification layer(s) to an already trained CNN, and then trains the network by using higher learning rates for the new layers, and lower learning rates for the pre-trained layers. This is to preserve the pre-trained parameters that are already relatively well generalized for the task and only make minor updates to adjust to the new classification objective. Thus, in this particular work, the classification layer of the Resnet-50, the fc1000 containing 1000 object classes from the ImageNet data in Fig. 4.13, was changed into a 12 object classes layer. The learning rate for this new layer is set to 0.001, while the remaining unchanged layers were set to have 0.0001 as the learning rate.

Using the data stated in Table 4.1, 400 images per class were set aside for the training, where 250 of them were used to fine-tune the model and 150 were used for validation. Hence, both the MSCOCO and ExDark provide 4,800 training images each, while the remaining 2,862 and 2,563 respectively make up the test sets. Experimentation using ratios from 10:0 (only bright images) to 0:10 (only low-light images) of bright to low-light images to fine-tune the model were conducted to observe the classification outcomes.

A few inferences can be drawn from the results shown in Table 4.3. First, the notion that the illumination variation of low-light can be addressed with the same manner

as noise (small additions into the training data) is improper. As shown in the results, the models that were fine-tuned with less amount of low-light images are weaker at classifying them, and gradually increases with the ratio. On the other hand, there was a presumption that balanced or generalized training data would enable the model to learn features that are mutually useful for both types of images and subsequently achieve best classification performance, but this is proven to be indiscreet. While the overall classification accuracy of Model 6 is the best, it appears to be a trade-off result as its performance is no better than a model specifically trained and tested on either bright (Model 1) or low-light (Model 11) images, even though they are addressing the same classification task. Hence, further analysis into the features was done to understand this behavior.

Feature Analysis with t-SNE. The features learned by the Resnet-50 model fine-tuned on 5:5 data ratio (Model 6) were analyzed using the t-SNE algorithm[4] (Maaten & Hinton (2008); Van Der Maaten (2014)). t-SNE is a dimensionality reduction technique for visualizing high-dimensional data in a 2D or 3D plot, referred to as embedding. It computes the pairwise similarities of the input data points, e.g. high-dimensional feature vectors, and the pairwise similarities of a corresponding low-dimensional counterpart, i.e. the 2D or 3D coordinates representing each vector. Then by minimizing the difference between these two distributions, the low-dimensional counterpart can sufficiently represent each high-dimensional datum in relation to other data and be plotted in a 2D or 3D space. Thus, it is noted that the resultant visualization by the t-SNE shows the relationship between each data instead of the absolute representation of a single datum in a lower dimension. However, this algorithm is computationally intensive, therefore the Principal Component Analysis (PCA)[5], a faster but less effective algorithm, is first used

[4]https://lvdmaaten.github.io/tsne/

[5]PCA is a dimension reduction technique implementing eigenvalues and eigenvectors to compute the correlation between data points to give the minimum variables while maintaining maximum variation that represents the original data.

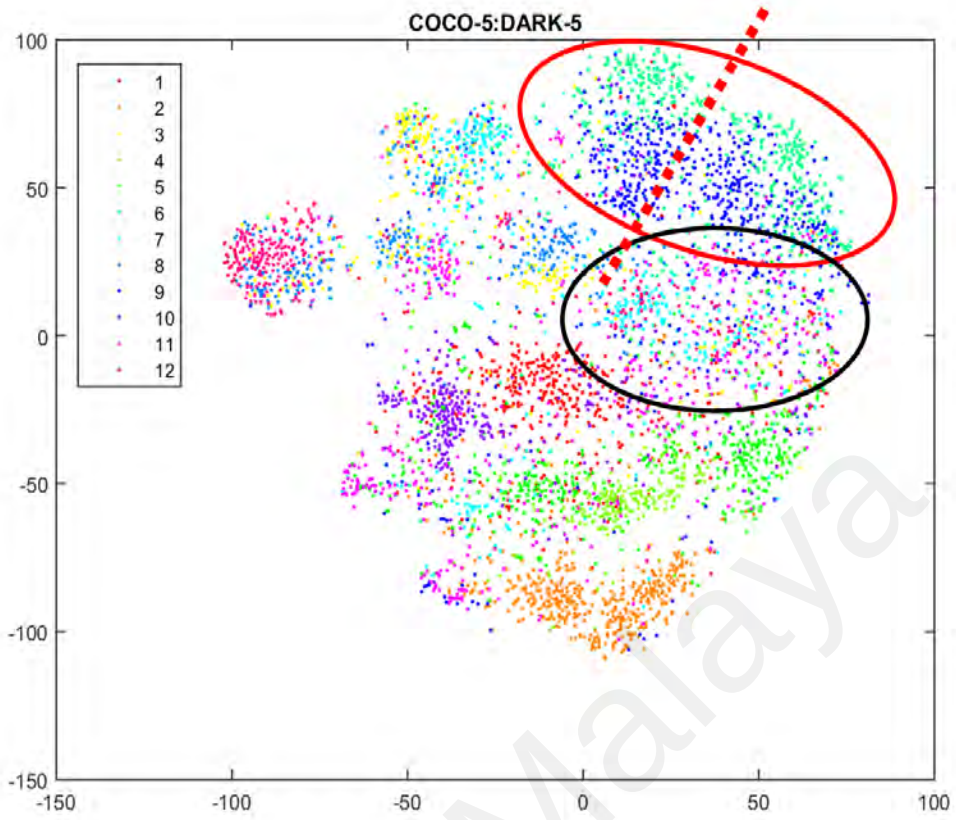


Figure 4.14: t-SNE embedding of features vectors from Resnet-50 fine-tuned on 5:5 ratio of bright and low-light images. Class 1-12: Bicycle, Boat, Bottle, Bus, Car, Cat, Chair, Cup, Dog, Motorbike, People, and Table.

to reduce the high-dimensional vector into an intermediate dimension, before the t-SNE further reduces it into the required dimension for embedding.

In the Resnet-50 model, the output produced by the last convolution layer is the high level representation that goes through a pooling layer that reduces its dimension and subsequently used by final fully connected layer for classification. Hence, to study the behavior of the high level features, the feature vectors produced by the last pooling layer of Model 6 when classifying the testing images were extracted. The PCA was used to first reduce these $1 \times 1 \times 2048$ dimension feature vectors into 50 dimensions and then the t-SNE further reduces it into a 2 dimension embedding which shows the relationship between the features.

Figures 4.14 and 4.15 show the embedding of the test images generated by t-SNE

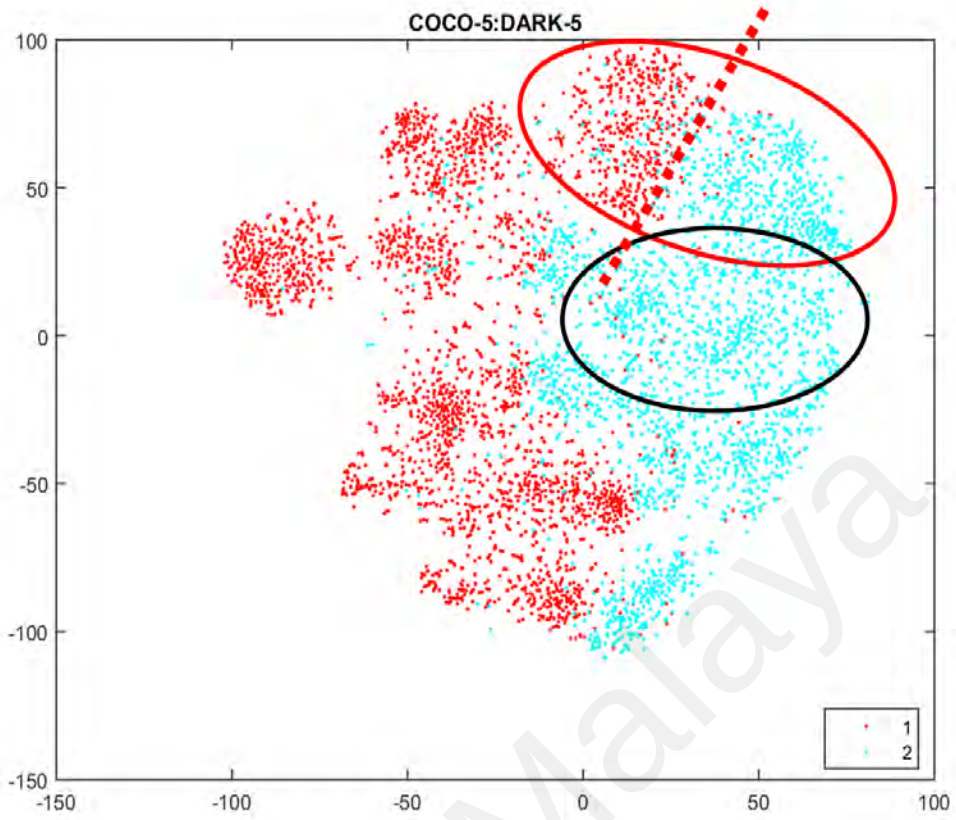


Figure 4.15: t-SNE embedding of features vectors from Resnet-50 fine-tuned on 5:5 ratio of bright and low-light images. Type 1-2: Bright (MSCOCO), and Low-light (ExDark) images.

and color coordinated by the object classes, and image types. Noticeable grouping of the object classes can be seen in Fig. 4.14, and classes that are relatively similar, such as Cat (5-green) and Dog (9-dark blue) are grouped closely as well, as circled in red. It is deduced that the learned features are able to capture high level abstraction of objects, though considerable amounts of confusion are still present, as circled in black.

A further study of feature embeddings from another perspective was done by differentiating the bright and low-light images, by marking the scatter points with two colors as in Fig. 4.15. Surprisingly, it shows a clear separation between bright images from the MSCOCO dataset (red) with the low-light images of ExDark (blue), which is a clear indication that even though the model is trained on both types of image for the same task, the features learned are inherently different. For example, the region for Cat and Dog

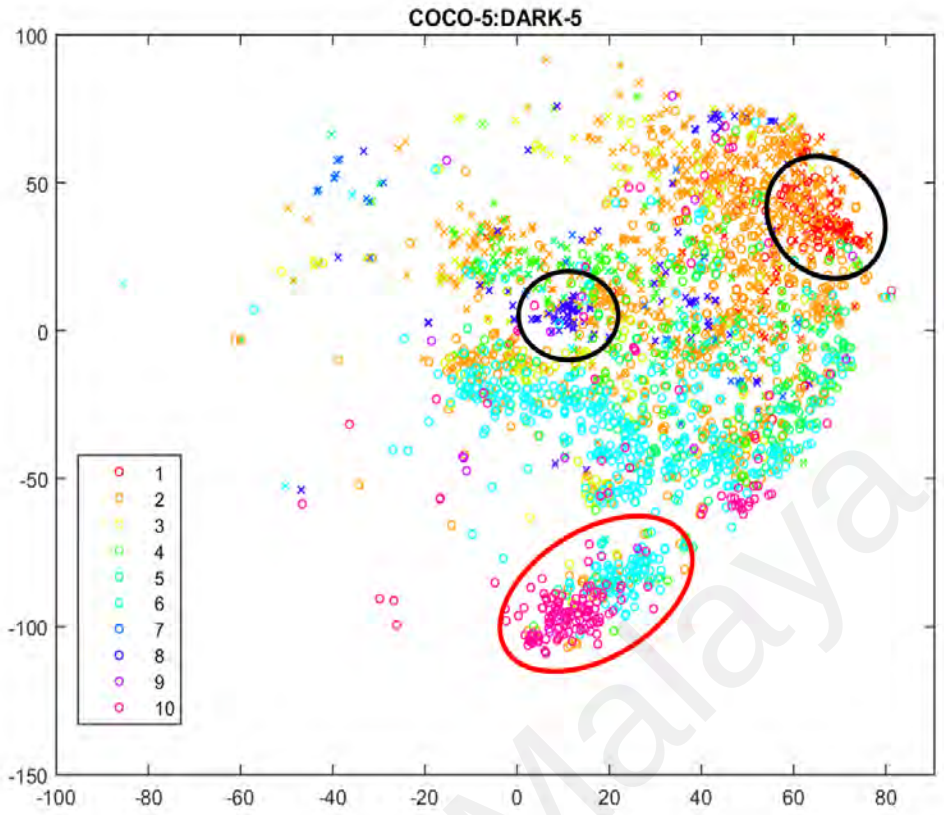


Figure 4.16: t-SNE embedding of features vectors from Resnet-50 fine-tuned on 5:5 ratio low-light images only. Separated by indoor ('x') and outdoor ('o') and color coded by the type of light conditions, 1-10: Low, Ambient, Object, Single, Weak, Strong, Screen (indoor only), Window (indoor only), Shadow (Outdoor only), and Twilight (outdoor only).

classes (circled in red) has a distinct split (red dotted line). Moreover, the region that do not have a distinct clustering of classes (circled in black) are found within the low-light image cluster, thus pointing out that the the features learned for low-light images maybe not be as robust as those for bright images.

Furthermore, the embedding was examined by color coordinating the scatter plot based on the types of low-light, as well as differentiating them by indoor and outdoor environments, as illustrated in Fig. 4.16 and 4.17. Firstly, the features seem to be able to distinguish indoor and outdoor by a small degree where the indoor images seem to cluster to the upper half of the embedding while the outdoor images are scattered throughout. On the other hand, the features appear to have the ability to distinguish certain types of low-

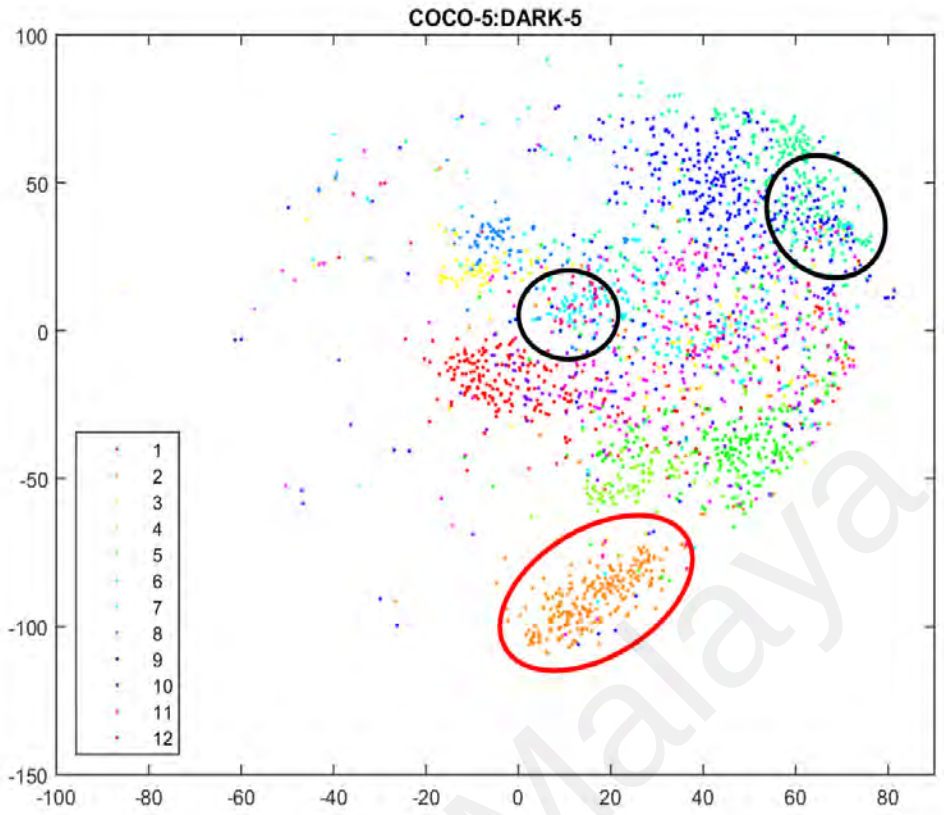


Figure 4.17: t-SNE embedding of features vectors from Resnet-50 fine-tuned on 5:5 ratio low-light images only. Color coded by classes, Class 1-12: Bicycle, Boat, Bottle, Bus, Car, Cat, Chair, Cup, Dog, Motorbike, People, and Table.

light images, such as Low (1-red), Strong (6-light blue), and Twilight (10-pink), though this ability may interfere with its robustness for the object classification task. As shown in the comparison between Fig. 4.16 and 4.17, the clustering of Low (1-red) and Window (8-dark blue) illumination type features (circled in black) has caused confusion to Cat, Chair, Dog, and People object classes. However, the clustering of the features may be stronger for the classification task, such as the Boat class cluster (circled in red) grouping both Strong and Twilight images together, though a separation can still be seen. Hence, it can be surmised that CNN model unwittingly learns low-light properties which can be a hindrance to the object classification task.

Attention Analysis with Activation Maps. This section delves into the activation maps of the trained model to find out its attention when performing the classification, and

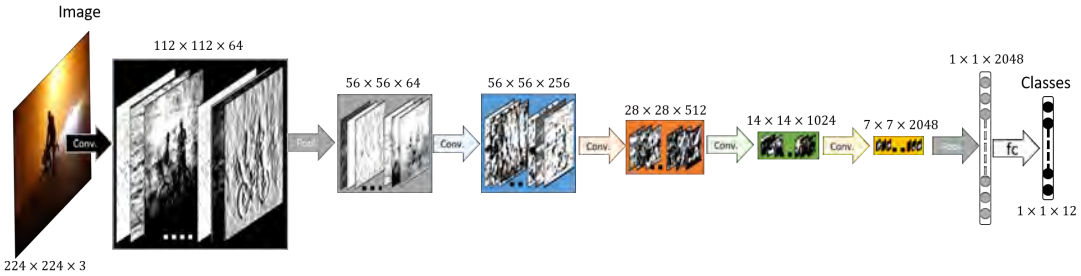


Figure 4.18: Feature maps produced by convolution and pooling layers in a simple CNN architecture. The dimensions of the feature maps get increasingly small while the number of maps increase as the more convolution and pooling operations are performed.

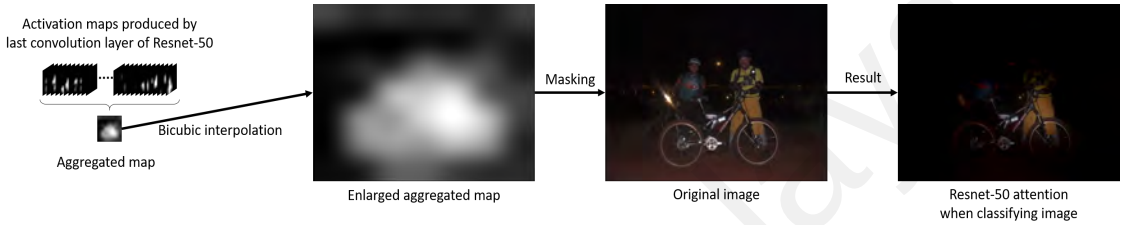


Figure 4.19: Visualization process for analyzing activation maps of Resnet-50.

if low-light elements are an influence to it. Activation maps are outputs produced by the convolution and pooling layers of a CNN architecture. Figure 4.18 shows the flow of a simple CNN and the feature maps produced after each operation, where it can be seen that the dimension of the maps decreases while the number of maps increases. Based on the findings of Yosinski et al. (2015), the early layers of a CNN captures lower level features such as edges and textures, and as the layers progress, increasingly high level concepts are encoded, which were then used for classification. Therefore, the crucial concepts for classification performance are found in the deeper layers, which is the target for this study.

Specifically, the activation maps before the last pooling layer of Model 6 (last convolution output before fully connected) was chosen to be visualized, so that the spatial location of the activations are preserved. The visualization process is illustrated in Fig. 4.19. First the $7 \times 7 \times 2048$ dimension activation maps of the Resnet-50 model when classifying an image were extracted. These maps are then aggregated into a single map by maximum pooling across the maps for every spatial location. Thus, the resultant ag-

(a) Bicycle (b) Bottle (c) Car (d) Boat (e) Bus

(f) Bottle (g) Cup (h) Bottle (i) Car (j) People

(k) Cat (l) Dog (m) Bus (n) Chair (o) Cat

Figure 4.20: Test images (top) and the visualization of activation maps (bottom). (a)-(e) Correctly classified low-light images; (f)-(j) Misclassified low-light images; (k)-(o) Misclassified bright images. (Classification results in sub-caption; correct class labels: (f) Cat, (g) Chair, (h) Cup, (i) Dog, (j) Motorbike, (k) Motorbike, (l) People, (m) Dog, (n) Table, (o) Bicycle).

gregated map will either have high values for locations that are highly activated or gives high contribution to the classifier. This map is then resized using bicubic interpolation to the original image's dimensions and superimpose onto the image, whereby the less infor-

mative regions will be masked and the remaining regions show the model's attention on the image that lead to the classification result.

Figure 4.20 shows a few examples of the classified test images and their respective activation regions. It is found that in the low-light images, the attention of the model are often drawn to the bright sources of light, either partially or entirely. For example, the activation maps of the correctly classified images in Fig. 4.20a - 4.20e show that while the main attention is on the object of interest, the light sources are either within the attention (Fig. 4.20a - 4.20c) or directly shined on the objects (Fig. 4.20d - 4.20e). While the model can "overlook" the light sources, like in Fig. 4.20e, there are many cases, such as Fig. 4.20f - 4.20j, where the attention of the model is overtaken by the brightest areas and causes misclassification. Yet this is not an issue for bright images, where the misclassification is commonly due to the attention being on another object instead of the labeled class as shown in Fig. 4.20k - 4.20o.

4.3 Summary

This chapter detailed the extensive analyses performed on low-light images from the perspective of both low and high level computer vision. Low level analysis looks into the pixel intensities of low-light images from global intensity histogram to local patch intensity variations. It is found that not only are the intensity patterns of low-light images greatly different from bright images, there exists subtle variations between the global intensity histograms of different types of low-light images. Moreover, the local patch intensity variations also greatly differ from one another due to the influence of light sources.

On the other hand, high level analysis digs deep into the behavior of common object features, both hand-crafted and learned, in which interesting insights were found. Foremost, the current design of hand-crafted features are mainly for bright conditions, thus unable to adequately address cases of noise and lack of details that frequently exist

in low-light images. Conversely, the investigation into learned features by visualizing the feature vectors and activation maps of a CNN, has lead to the understanding that low-light "alters" object features, i.e. the same object in bright and low-light yields amply different features. Moreover, the irregularity of illumination greatly challenges the attention of features that is not found in bright environments. Therefore, the low-light phenomenon in computer vision is not to be trifled with lightly, but instead requires careful consideration.

CHAPTER 5: LOW-LIGHT IMAGE CONTRAST ENHANCEMENT USING GAUSSIAN PROCESS

Low-light is a condition that challenges computer vision systems that extends beyond the common conception of illumination invariant features as discussed in Chapter 4. Thus, instead of looking into feature designs and model optimizations, this research addresses the challenges through the perspective of features retrieval via contrast enhancement.

Noting the global and local illumination variation, this chapter details the proposed approach that is unlike previous works. Inspired by the localized and spatially non-linear nature of human eyes in adaptation (Rose (1948); Vangorp et al. (2015)), the aim is to develop a locally adaptable model to enhance the contrast of a single low-light image[1] with emphasis on retrieving informative details instead of aesthetic restoration. Specifically, to enhance low-light images using localized functions exclusive to individual regions or pixels, thus, distinct enhancements are assigned to different illuminations for optimal results. To this end, the $\mathcal{GP}$ regression is employed to construct a distribution of functions for a precise enhancement due to its sophistication and robustness on such localized data.

5.1 Problem Formulation

Formally, the global and local illumination variation of low-light images are a result of real world behavior of light defined by both the light source and the inverse-square law of distance:

$$E(p) = \frac{L(p)\cos^2\theta dA}{l^2} \tag{5.1}$$

where E is the irradiance or intensity per unit area on the object, L is the radiance from light source, $\cos\theta$ is the foreshortening, dA is the unit area, and l is the distance between

[1]In this research, low-light images are the primary target for the contrast enhancement, instead of low contrast images which can be either low-light or bright images.

the light source and object. Due to this, the light intensity of an object depends on radiance of the light source (e.g. midday, twilight, etc.) which is the global variation, and the foreshortening and distance are the causes of local variation. It is noted that real bright images also have the foreshortening effect caused by local illuminance variations, and coincidentally brings out the details of contents within an image, therefore, considering the third objective of this work, this characteristic is preserved instead of brightening every dark region in low-light images.

Thus, the low-light image contrast enhancement is essentially improving the remaining of the scene illuminance that was captured. Figuratively, this is achieved by reversing the effect caused by largely varied light sources L and distances l on the intensity. Hence, the low-light image enhancement can be modeled as:

$$I_B(x) = I_D(x)\mathcal{F}(L, l_x), \tag{5.2}$$
$$\text{s.t.} \quad L = \{L_1, L_2, ..., L_i\}; l_x = \{l_{x_1}, l_{x_2}, ..., l_{x_j}\}$$

where $I_B(x)$ is the enhanced image that has an appropriate contrast and relative uniform intensity distribution, $I_D(x)$ is the captured low-light image that has relatively low illuminance, low contrast and intensity variation, x is a pixel or small patch in the image, and $\mathcal{F}(L, l_x)$ is the mapping operator defined by the light source L and distance l. However, estimating the mapping operator is not a simple task due to the following reasons:

- Diverse types of light sources L have individual light strength providing distinct levels of intensity;
- Multiple sources of light $\{L_1, L_2, \cdots, L_i\}$ increase the non-uniformity of the scene luminance;
- Any point of a scene can locate at a varying distance $\{l_{x_1}, l_{x_2}, ..., l_{x_j}\}$ between the

object and source of light.

Therefore, Equation 5.2 is transformed into:

$$I_B(x) = f(I_D(x)) \tag{5.3}$$

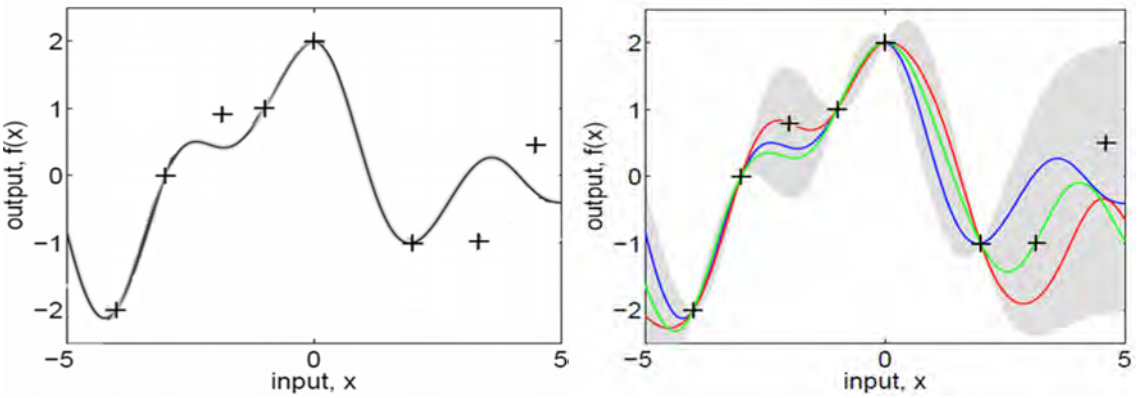
where $f(\cdot)$ is the enhancement function that models the relationship of pixels in the low-light image, $I_D(x)$ with the target bright pixels of $I_B(x)$.

5.2 Proposed Method

To reiterate, the relationship of pixels between $I_D(x)$ and $I_B(x)$ are localized, hence $f(\cdot)$ is not a single function acting on all pixels, but a collection of different functions that act on the respective pixels. Therefore, $\mathcal{GP}$ is the reasonable framework for modeling such relationships as it defines a distribution over functions.

5.2.1 Gaussian Process Overview

Williams & Rasmussen (2006) define the $\mathcal{GP}$ as *a collection of random variables where any finite number of which have (consistent) joint Gaussian distributions*, i.e. the distribution of a $\mathcal{GP}$ is the combined distribution of random variables. It is inherently different from Gaussian Distribution (GD) such that, in a GD, individual random variables or inputs x in a vector are indexed by their position in the vector such that the distribution is defined by the mean, μ which is a vector, and covariance, $\sum$ that is matrix. The $\mathcal{GP}$ on the other hand, has the input x as the index associated with the random variable $g(x)$ which itself is a distribution, and the $\mathcal{GP}$ mean, m and covariance, k are functions. Simply put, the GD is a distribution over vectors whereas $\mathcal{GP}$ is a distribution over functions as illustrated in Fig. 5.1. The GD plots a single function that best fits the given input and



(a) Gaussian Distribution **(b) Gaussian Process**

Figure 5.1: The GD is a single function that best fits the given data, whereas the $\mathcal{GP}$ consists multiple functions (gray area) that are shaped by the data.

output data, whereas the $\mathcal{GP}$ is made up of multiple functions, shown as the gray area, that are shaped by the given data.

Mathematically, the $\mathcal{GP}$ defines a distribution over function f that estimates an output y from the marginal distribution of functions $\mathbb{P}(f(x_1), f(x_2), \cdots, f(x_k))$ of finite inputs $x = \{x_1, x_2, \cdots, x_k\}$. It is parameterized by a mean function $m(x)$ and a covariance function $k(x_{tr}, x_{ts})$ such that $f(x) \sim \mathcal{GP}(m(x), k(x_{tr}, x_{ts}))$, where the joint distribution of training and test outputs is:

$$\begin{bmatrix} y_{tr} \\ y_{ts} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_{tr}) \\ m(x_{ts}) \end{bmatrix}, \begin{bmatrix} K(x_{tr}, x_{tr}) & K(x_{tr}, x_{ts}) \\ K(x_{ts}, x_{tr}) & K(x_{ts}, x_{ts}) \end{bmatrix} \right). \tag{5.4}$$

y_{tr} refers to the training outputs and y_{ts} is the testing outputs in which $y = [y_{tr}, y_{ts}]$. Similarly, $x = [x_{tr}, x_{ts}]$ are the training and testing inputs. $K(x_{tr}, x_{ts})$ denotes the covariance matrix between the training and testing inputs, along with $K(x_{tr}, x_{tr})$,$K(x_{ts}, x_{tr})$, and $K(x_{ts}, x_{ts})$ as the covariance of their respective pairings. Given the observation y_{tr} from

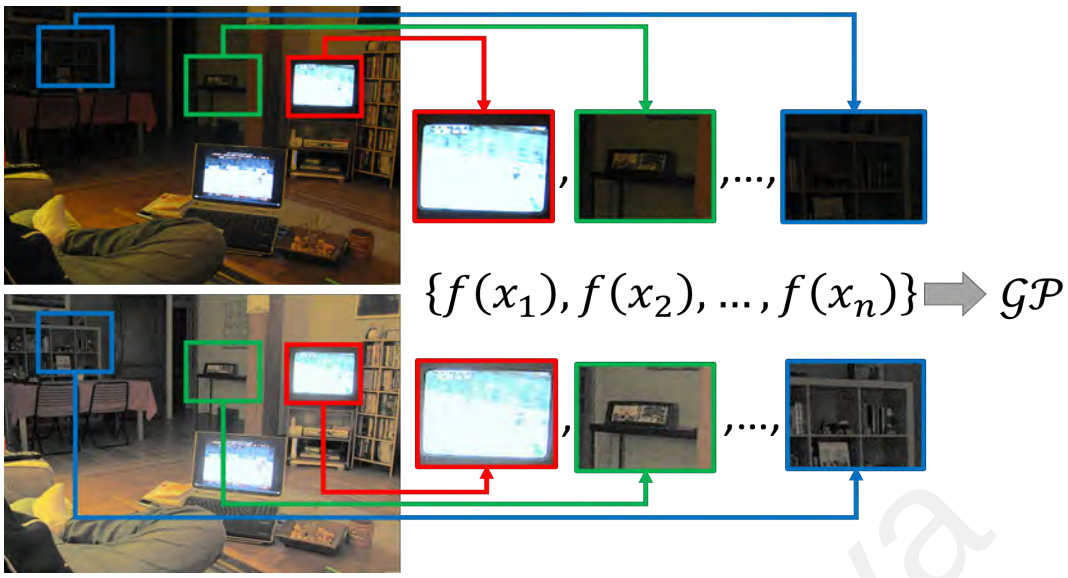


Figure 5.2: The intuition is to have localized enhancement functions for each region/pixel, the $\mathcal{GP}$ is used to govern them into a distribution of functions. (Top: Low-light image, Bottom: Contrast enhanced image).

x_{tr}, the output y_{ts} can be estimated with x_{ts} from the conditional distribution:

$$\mathbb{P}(y_{ts}|x_{tr}, y_{tr}, x_{ts}) \sim \mathcal{N}(\mu, \Sigma) \tag{5.5}$$

$$\text{where} \quad \mu = m(x_{ts}) + K(x_{ts}, x_{tr})[K(x_{tr}, x_{tr})]^{-1}(y_{tr} - m(x_{tr})), \tag{5.6}$$

$$\Sigma = K(x_{ts}, x_{ts}) - K(x_{ts}, x_{tr})[K(x_{tr}, x_{tr})]^{-1}K(x_{tr}, x_{ts}) \tag{5.7}$$

5.2.2 Modeling Contrast Enhancement with Gaussian Process

The objective is to estimate a corresponding bright image I_B given a single low-light image I_D. As previously established, the enhancement functions for low-light images are localized, thus the distribution of functions $\mathbb{P}(f(x_1), f(x_2), \cdots, f(x_k))$ are therefore the varied local luminance mapping functions as shown in Fig. 5.2, where the inputs x refer to either the local patches or pixels in the low-light image. Specifically, the implemented testing inputs are the pixels from the low-light image, i.e. $x_{ts} = \{p_D | p_D \in I_D\}$, while the testing outputs are the corresponding enhanced image pixels, i.e. $y_{ts} = \{p_B | p_B \in I_B\}$. Figure 1.5 shows the overall flow of the proposed enhancement framework.

The construction of the $\mathcal{GP}$ is specified by mean and covariance functions given the data. Hence, for these priors, the zero mean function $m(x) = 0$ is used to simplify the modeling process and allow the relationship between x_{tr} and x_{ts} to be fully defined by the covariance function $k(x_{tr}, x_{ts})$. The squared exponential is chosen as the covariance function:

$$k(x_{tr}, x_{ts}) = \sigma_f^2 \exp\left(-\frac{(x_{tr} - x_{ts})^2}{2d^2}\right), \tag{5.8}$$

where hyperparameters, σ_f^2 is the data variance, and d is the length scale that defines the smoothness of the $\mathcal{GP}$. These hyperparameters $\theta_{\mathcal{GP}} = \{\sigma_f, d\}$ determines the form of the distribution of functions, and are inferred from the low-light data using conjugate gradients to optimize the log marginal likelihood:

$$\mathcal{L} = \log \mathbb{P}(y_{tr} | x_{tr}, \theta_{\mathcal{GP}}). \tag{5.9}$$

As the posterior distribution is data dependent, each image is therefore enhanced by image exclusive optimal hyperparameters in the framework. For the training data, the patches of $m \times n$ pixels of the given low-light image I_D are defined as the training inputs x_{tr} whereas the training outputs y_{tr} are patches of the similar pixels dimension and spatial location from a corresponding bright image I_E, as illustrated in Fig. 1.5. To this end, the training inputs $\bar{P}_D = \{\bar{P}_{D,1}, \bar{P}_{D,2}, \cdots, \bar{P}_{D,k}\}$ and outputs $\bar{P}_E = \{\bar{P}_{E,1}, \bar{P}_{E,2}, \cdots, \bar{P}_{E,k}\}$ are the average intensities of local patches, $P_D = \{P_{D,1}, P_{D,2}, \cdots, P_{D,k} | P_D \subseteq I_D\}$ and $P_E = \{P_{E,1}, P_{E,2}, \cdots, P_{E,k} | P_E \subseteq I_E\}$.

From a preliminary investigation, it was found that the sizes of P_D and P_E heavily influence the posterior distribution and quality of the enhancement, where smaller sizes give better results as shown in Fig. 5.3. This is because the region size constrains the precision of the mapping distribution. As the size decreases, more constraints are introduced

Figure 5.3: Low-light image and enhancement results using $\mathcal{GP}$s trained by different patch sizes. From left: Original low-light image, results using patch sizes 4×4, 8×8, 16×16, and 32×32.

and brings more precise distribution. However, there is a trade-off where more constraints will drastically increase the computational cost. Due to this reason, an inflection point is found where the P_D and P_E are fixed as patches of 4×4 pixels. In addition, many patches in an image bore a superficial resemblance, therefore the computation is further optimized by using only one region-pair instead of multiple similar ones as shown in Fig. 5.4, i.e.

$$\begin{aligned} \{x_{tr}, y_{tr}\} &= \{P_{D,i}, P_{E,i}\} \\ \text{if} \quad \left\{\bar{P}_{D,i} \approx \bar{P}_{D,j} | \left\{\bar{P}_{D,i}, \bar{P}_{D,j}\right\} \subset \bar{P}_D\right\} & \\ \vee \left\{\bar{P}_{E,i} \approx \bar{P}_{E,j} | \left\{\bar{P}_{E,i}, \bar{P}_{E,j}\right\} \subset \bar{P}_E\right\} & \\ \text{where} \quad i \neq j & \end{aligned} \tag{5.10}$$

The average intensity of the patches are used as the similarity measure.

5.2.3 Gaussian Process Training Data Estimation

In order to build the $\mathcal{GP}$, one will require the input and output training data, akin to building any distribution functions as explained in Section 5.2.1 and illustrated in Fig. 5.1. For example, He et al. (2011) adopted the $\mathcal{GP}$ for single image super-resolution, where training input and output were generated using upsampling and downsampling operations.

In the low-light image enhancement task, the input data is the low-light image I_D, whereas the desired output data is the corresponding bright image. However, this image is

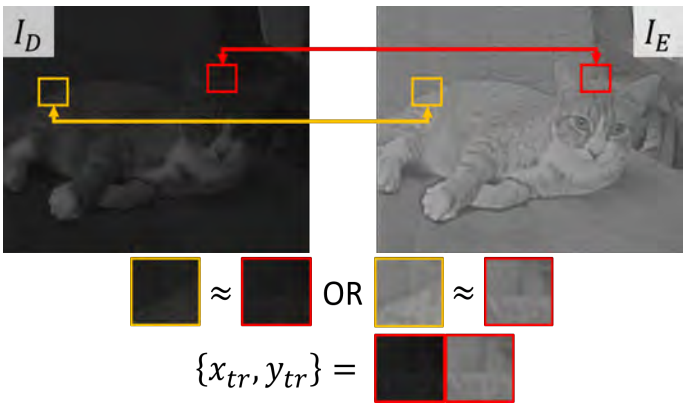


Figure 5.4: $\mathcal{GP}$ training input I_D (in luminance channel, Y) and output I_E, and training data optimization. Note that the training data x_{tr} and y_{tr} are corresponding patches from I_D and I_E respectively, enabling edge/texture relationships to be preserved in the $\mathcal{GP}$. If there are multiple patches pairs that are similar, only one pair is used for training to minimize computational cost.

not readily available because the most precise reference would be the bright counterpart of the exact same scene in order for specific relationships between the low-light and bright pixels to be formed. This relationship is not transferable from the distribution of other images because the texture and edge details within an image are different from another image. Hence, as shown in Fig. 1.5, a CNN is employed as an intermediate transformation model to generate the training output I_E. This I_E is also an enhancement of I_D, but the difference of it with I_B is that it is not locally optimized. Even so, it is globally optimized by the trained CNN on large low-light data and maintains sufficient details and textures from I_D, as shown in Fig. 5.4, to build the $\mathcal{GP}$. The choice of CNN as the intermediate model is due to the successful results of pixel-wise transformation works such as image denoising (Jain & Seung (2009)) and super-resolution (C. Dong et al. (2015)). While these are different domains of image enhancement, it shows a promising solution as a subcomponent of the proposed framework.

The more profound reason is that low-light image enhancement is also considered as a sparse coding problem where features such as textures and edges should be improved in addition to brightening (Loh & Chan (2015)). The sparse representations of low-light images and bright images are to be coded in dictionaries and a solver optimizes the non-

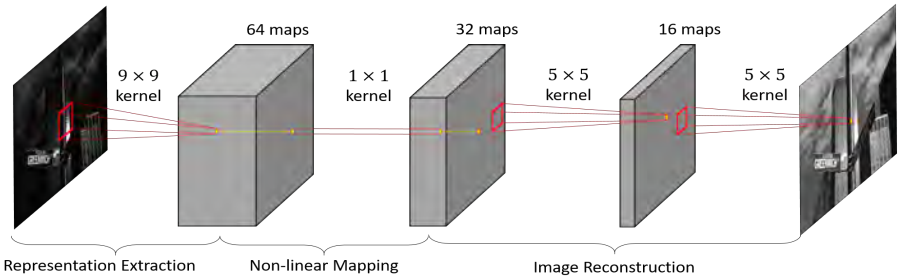


Figure 5.5: CNN architecture modified from C. Dong et al. (2015)'s model, that is used to generate the training output for $\mathcal{GP}$.

linear mapping between these dictionaries. Representations of any low-light image are projected onto the low-light dictionary, mapped to the bright dictionary, and subsequently reconstructed into a bright image. It is made clear by C. Dong et al. (2015) that CNN is a well optimized variant of sparse coding due to its exhaustive end-to-end optimization that includes image representations, dictionaries and non-linear mapping operations. Hence, their model's architecture is modified by incorporating an additional layer, as shown in Fig. 5.5, to increase network complexity for producing the training output of $\mathcal{GP}$.

Let $g : I_D \mapsto I_E$ represent the mapping operation from the low-light image I_D to the reference image I_E. The network consists of four convolution layers serving as representation extraction, non-linear mapping, and image reconstruction operators as shown in Fig. 5.5. The first layer acts as the patch based representation extractor with the following operation:

$$g_1(I_D) = \max(0, w_1 * I_D + b_1), \tag{5.11}$$

where b_1 is the bias and w_1 is the learned feature extraction filters of size $9 \times 9 \times c$. This operation is likened to local features extraction such as SIFT and HOG in the sparse coding approach, but extracts optimal low-light representation instead of predefined estimation of features. The c parameter refers to the number of channels in I_D which can, but not limited to be, 1 if it is the luminance channel or 3 for RGB color space. The second

convolution performs the non-linear mapping operation:

$$g_2(I_D) = \max(0, w_2 * g_1(I_D) + b_2). \tag{5.12}$$

The bias, b_2 and filters, w_2 learn the mapping relationship between the low-light representation and bright representation, similar to the sparse coding solver. The last two convolution layer acts as the image reconstruction module:

$$g_3(I_D) = \max(0, w_3 * g_2(I_D) + b_3). \tag{5.13}$$

$$I_E = g(I_D) = w_4 * g_3(I_D) + b_4. \tag{5.14}$$

The filters w_3, w_4 and biases b_3, b_4 in these layers project the bright representations back to the image space similar to an inversion function that converts feature vectors back into images. This double layer reconstruction is used because inverting features back into images is an innately more complicated problem as compared to feature extraction, hence requiring more non-linearity to address. The final image produced is the I_E which has the same c channels as I_D.

5.2.3 (a) Loss Function

The training of the CNN model optimizes the parameters,
$\theta_{\text{CNN}} = \{w_1, w_2, w_3, w_4, b_1, b_2, b_3, b_4\}$ by minimizing a loss function, $\mathcal{L}$. In this work, the loss function used is the Mean Squared Error (MSE):

$$\mathcal{L}(I_E, I_B) = \frac{1}{n} \sum_{i=1}^{n} \|I_E - I_B\|^2 \tag{5.15}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \|g(I_D|\theta_{\text{CNN}}) - I_B\|^2, \tag{5.16}$$

where n refers to the number of training data. This function is minimized by stochastic gradient descent and backpropagation with the following weight and bias updates for each layer:

$$w_i^{\text{new}} = w_i^{\text{old}} - \eta_i \frac{\partial g}{\partial w}(w_i^{\text{old}}), b_i^{\text{new}} = b_i^{\text{old}} - \eta_i \frac{\partial g}{\partial b}(b_i^{\text{old}}) \tag{5.17}$$

where $i = \{1,2,3,4\}$, and η_i is the learning rate for the convolution layers. The values are set as $\eta_1 = \eta_2 = 10^{-4}$, and $\eta_3 = \eta_4 = 10^{-5}$. Minimizing MSE updates the weight to produce the reconstruction I_E that possess high PSNR (C. Dong et al. (2015)), the commonly used metric for image quality evaluation, which in turn setups a output training data with sufficient quality for the $\mathcal{GP}$. However, it is found that PSNR may not be the ideal image quality measure in this problem, as detailed in Section 5.3.4.

5.2.3 (b) Transformation Model Training

Up till this point, there is no dataset of bright and low-light image pairs available publicly that is large enough to sufficiently generalize a model, and it is impractical to capture the pairs. Therefore, synthetic darkening of real bright images I_b is engaged to generate the artificial low-light counterpart I_d and subsequently use them as training pairs.

The darkening operation used is a combination of contrast scaling and gamma correction:

$$I_d = C_{lim} I_b^{\gamma}, \tag{5.18}$$

where C_{lim} is the upper intensity limit of I_d and γ is the gamma value. Different combinations of C_{lim} and γ are applied to generate different levels of low-light images to be learned. Section 5.3.1 (a) details the justification on using this scheme to generate data that statistically approximates real low-light images.

While synthetic, the approximation towards real data is sufficient to train a reliable transformation model and more importantly provides large enough data for the model to

be optimized globally across many variations of low-light conditions. Subsequently, the model produces reliable training data for the $\mathcal{GP}$.

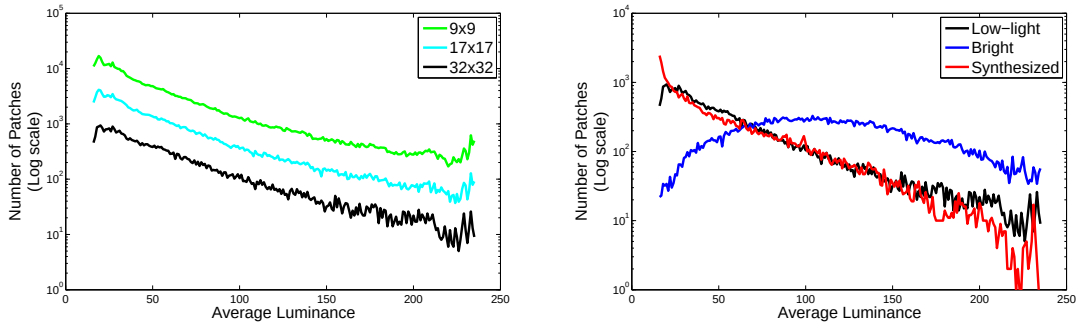
5.3 Experiments

This section describes the implementation details of the proposed method, including training data generation, and evaluation results in comparison to the latest work in low-light image contrast enhancement. The conventional quality metric, i.e. PSNR as well as a newly proposed information retrieval metric were employed to validate the proposed method. In order to standardize the experimental data for quantitative evaluations, images are again sub-sampled from the validation set of MSCOCO to obtain 300 bright images which were darkened according to the scheme stated in Section 5.2.3 (b). This gives a total of 7,500 testing images, a similar amount to the analysis set in Chapter 4, for quantitative evaluation. Additionally, 150 real low-light images were sampled from the MSCOCO dataset as well for analysis in Section 5.3.1 (a) and qualitative assessment.

5.3.1 Implementation Details

Generally, the proposed method can be applied to images in RGB color space, grayscale, or even luminance (Y) channel. In this experimental implementation, the luminance channel Y is used whereas the chrominance components (C_bC_r) are unaltered and only used for producing the final colored image. Similarly, the reference image estimated by the CNN model is in the Y channel.

Two variants of the CNN model were trained for comparison. The first is trained using the bright I_b and darkened I_d images pairs resized to $p \times q$ pixels, referred as CNN1. The other, CNN2 is trained using $m \times n$ pixels patches obtained by dividing the images I_b and I_d without any other spatial modifications.



(a) **(b)**

Figure 5.6: Distribution comparison of average intensities of image patches. (a) Between different patch sizes from the real low-light images only. (b) Between patches from both the real and synthesized low-light images with patch size of 32×32.

5.3.1 (a) CNN Training Data Generation

Firstly, in the attempt to closely simulate the darkness conditions of real images to train a reliable model to produce the $\mathcal{GP}$ training data, the intensity distributions of real low-light images were to be studied. Hence, 150 of such images were extracted from the MSCOCO dataset (non-overlapping with the testing images used in evaluations), inclusive of both indoor and outdoor environments, for analysis. These images were divided into non-overlapping patches and the average luminance of each patch was obtained. These values were then binned and the distribution trend was observed for patch sizes of $9 \times 9, 17 \times 17$ and 32×32. As shown in Fig. 5.6a, the number of patches extracted in the real low-light images logarithmically decrease as the intensity level increases, irrespective of the patch sizes.

Based on this observation, it is expected that by synthetically darkening bright images using Equation 5.18, the low-light image patches could exhibit a similar trend. To this end, 150 bright images were randomly sampled from the MSCOCO dataset, where Fig. 5.6b shows that the distribution of average luminance in their patches (blue) greatly differs from the low-light images (black). These bright images were then darkened with the combination of $C_{lim} = \{250, 200, 150, 100, 50\}$ and $\gamma = \{1, 2, 3, 4, 5\}$ to produce 25

(a) **(b)** **(c)** **(d)** **(e)**

Figure 5.7: Examples of low-light images synthesized from one bright image using different configurations. **(a)** $C_{lim} = 250, \gamma = 1$**; (b)** $C_{lim} = 200, \gamma = 2$**; (c)** $C_{lim} = 150, \gamma = 3$**; (d)** $C_{lim} = 100, \gamma = 4$**; (e)** $C_{lim} = 50, \gamma = 5$**.**

levels of darkening for a single bright patch. Figure 5.7 shows examples of different darkening levels of the same image. The intensity distribution of the synthesized patches of size 32×32 shows a similar trend (red) as that of real low-light image patches in Fig. 5.6b, particularly for lower intensity levels. Hence, the CNN training pairs were synthesized in this manner.

To build the training data, the entire MSCOCO training set which contains 82,783 images was used for the synthesis in order for good model generalization. Each image provides 26 training pairs (including the original bright patch paired with itself), hence providing a total of 2,152,358 images for CNN training and validation. For the CNN1 model trained using the full images, the data were resized to 256×256 pixels and normalized to the range of $[0, 1]$. Whereas, the training data of CNN2 are 32×32 pixels non-overlapping patches that were extracted from the same image set and similarly normalized. Additionally, the patches with average intensity of 0 were eliminated from the training set as such patches have no contrast and moreover, it is illogical to train the CNN

to predict details from a blank patch for an enhancement task.

5.3.2 State-of-the-art Methods

The proposed method is compared to the state-of-the-arts in low-light image contrast enhancement, which includes Luminance Adaptive Contrast Enhancement (LACE) (L. Li et al. (2015)), Fushion-Based Enhancing (FBE) (X. Fu, Zeng, Huang, Liao, et al. (2016)), Weighted Variational Model (WVM) (X. Fu, Zeng, Huang, Zhang, & Ding (2016)), and Low-light IMage Enhancement (LIME) (Guo et al. (2017)). Following are the descriptions of these methods:

- **LACE**, adopts the dark channel prior haze removal method by He et al. (2011) to perform low-light image contrast enhancement. He et al. (2011) proposed the hazy image model, $H = t \cdot J + (1 - t) \cdot A$ where H is the hazy image, t is the light transmission map, J is the restored haze-free image, A is the global atmospheric light value, and $(\cdot)$ is the element-wise multiplication. This model can be applied to low-light image enhancement based on the findings by X. Dong et al. (2011), where an inverted low-light image exhibits characteristics of a hazy image. The LACE applies the dehazing model on the inverted low-light image using an adaptive weight coefficient to estimate the t and then inverting the resultant J to obtain the enhanced low-light image. The full proposed method by L. Li et al. (2015) includes a prior denoising module, however, this component is not included in the experimental comparison of this research in order to establish a fair comparison with other methods that do no explicitly deal with noise. The code of LACE was reimplemented from the dark channel haze removal work by He et al. (2011)[2] based on the details given by the paper for the experiments in Sections 5.3.3 and 5.3.4.

[2]https://github.com/sjtrny/Dark-Channel-Haze-Removal

- **FBE** is a Retinex-based method where a low-light image is decomposed into illumination and reflectance as $S^c = R^c \cdot I$ where S is low-light image, R is the reflectance, I is the illumination, and c is the RGB color space. This method is executed in four steps, where firstly the illumination is estimated by taking the maximum values among the RGB channels, inspired by the dark channel prior for dehazing (He et al. (2011)), followed by smoothening and refinement. Step two derives three inputs from the obtained illumination: (1) the illumination itself, (2) the illumination augmented using arc tangent transformation for improved global luminance, and (3) the equalized illumination using CLAHE for enhanced local contrast. The next step involves computing pixel-level weights that determine the fusion of the three inputs based on the region quality of each input. Finally, a multi-scale fusion approach is implemented, where the three derived inputs are decomposed into Laplacian pyramids, while the generated weights are decomposed into Gaussian pyramids for fusion and then multiplying the respective levels of the pyramids and then upsampling for summation into a final enhanced illumination. This enhanced illumination is combined with the reflectance to obtain the enhanced low-light image. For evaluation and comparison, the code for this FBE[3] was obtained from the author for implementation.

- **WVM** is also a Retinex-type solution for low-light image enhancement. Its difference with FBE is that the WVM directly addresses the reflectance and illumination estimation problem instead of just computing the R from an estimated I by $R = \frac{S}{L}$. This approach transforms the model into the logarithmic domain and solves an objective function that outputs both the illumination and the reflectance. By estimating the reflectance in such a way, details are effectively preserved and error

[3]http://smartdsp.xmu.edu.cn/weak-illumination.html

propagation from the illumination is avoided. With these estimated components, the enhancement is proceeded by adjusting the illumination using gamma correction with empirically determined gamma parameter. The implementation of this WVM algorithm[4] is also provided by the author.

- **LIME** is a method built upon the Retinex model but takes an untraditional approach for the enhancement. This method uses a similar model, $L = R \cdot T$, however instead of decomposing the image L into reflectance and illumination, the model defines R as the desired image to be recovered and T as the illumination map. The T is therefore a sort of transformation map that induces the low-light element into images and thus effectively simplifying the problem into estimating only a T instead of two components. Similar to FBE, taking inspiration from the dehazing algorithm by He et al. (2011), the T is first estimated using the maximum values among the RGB channels. The map is then optimized using structure-aware prior to obtain the final illumination map. Additionally, the recovered image R is denoised to improve final image quality. Similar to the LACE, this component in the author provided code for LIME[5] is removed from the algorithm for fair comparisons in Sections 5.3.3 and 5.3.4.

5.3.3 Qualitative Evaluation

To demonstrate the significance of local luminance variation in this problem, the results generated by the differently trained globally optimized CNN models (CNN1 and CNN2) and their effect as the training output for the proposed locally optimized $\mathcal{GP}$ model ($\mathcal{GP}$(prior1) and $\mathcal{GP}$(prior2) respectively) were first compared. Figures 5.8 and 5.9 show examples of the models applied on real and synthesized low-light images respectively.

[4]http://smartdsp.xmu.edu.cn/cvpr2016.html

[5]https://sites.google.com/view/xjguo/lime

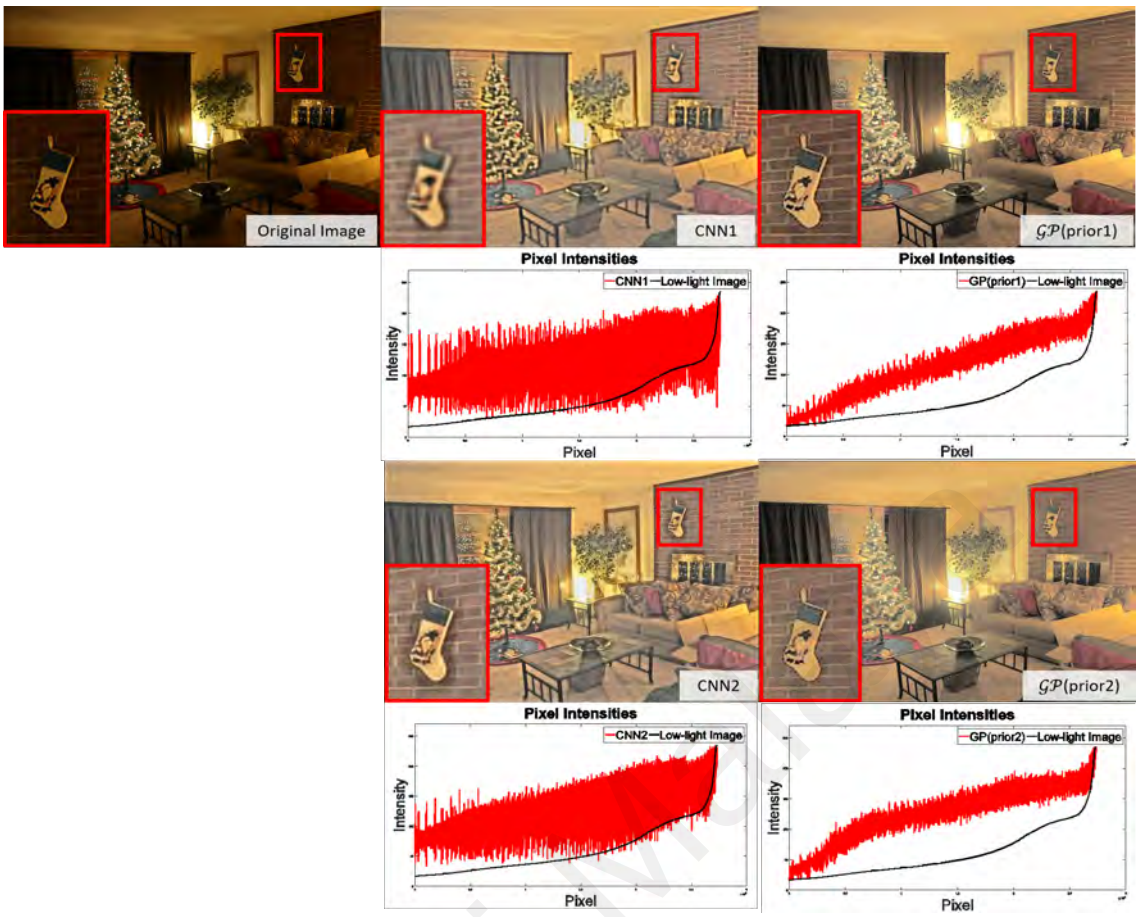


Figure 5.8: Example of the contrast enhancement on a real low-light image using 2 variants of the CNN and the proposed method, and the intensity of each pixel before and after enhancement (arranged in ascending order of pixel values from the original image).

Judging from their appearance, the CNN results look more artificial than $\mathcal{GP}$, mainly due to the difference of globally and locally optimized modeling. Further inspection shows that CNN1 reproduces better overall brightness, but CNN2 preserves more details than CNN1 which looks blur. This could be due to the lost of details when down-sampling the training images[6] to lower resolution for the CNN1.

While both of the reference images display evident differences, it is less apparent after the $\mathcal{GP}$. The result of real low-light images using $\mathcal{GP}$(prior1) has slightly better contrast due to its bigger perception whereas $\mathcal{GP}$(prior2) appears brighter, while there are not much differences for synthesized low-light images. However, it is clear in the pixel intensity distributions, that the CNN produces a scattered distribution whereas $\mathcal{GP}$

[6]It is possible to use original images for training, but the computational cost is impractical.

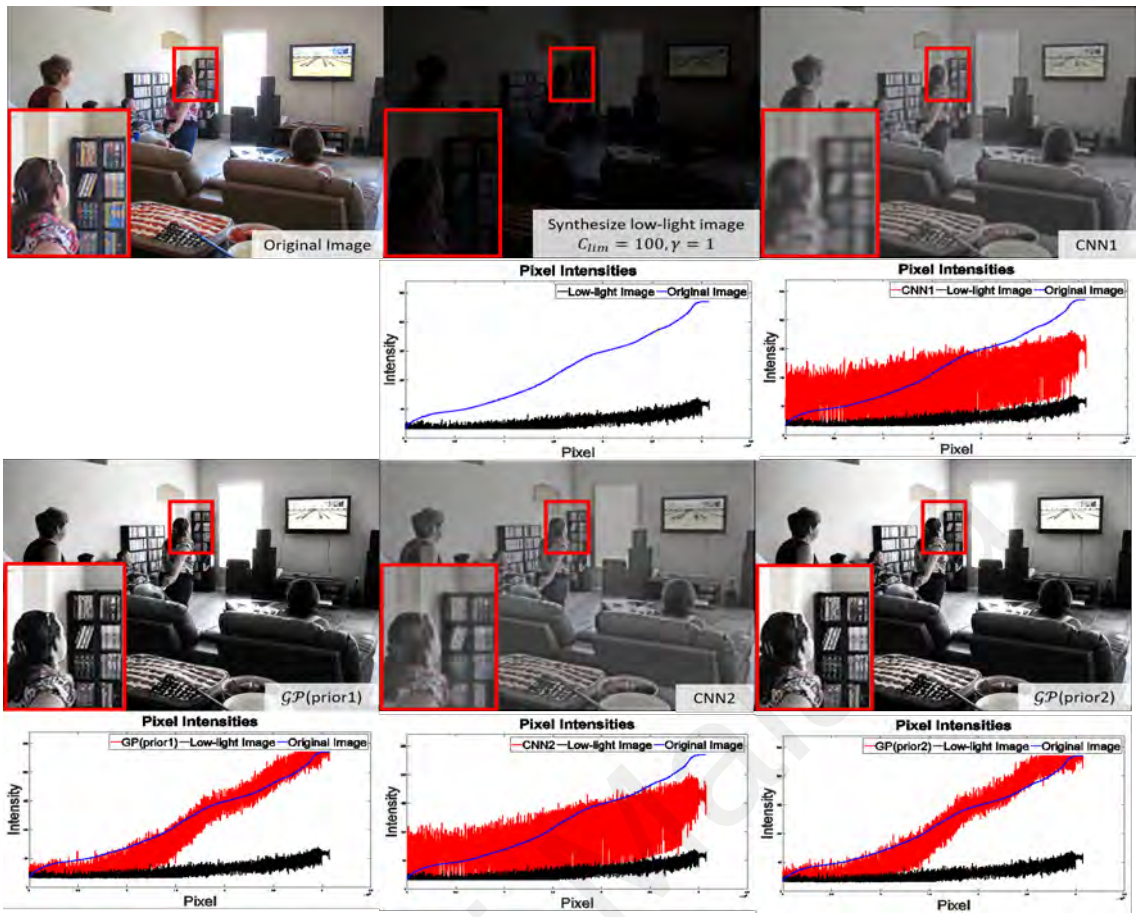


Figure 5.9: Example of the contrast enhancement on a synthesized low-light image using 2 variants of the CNN and the proposed model, and the intensity of each pixel before and after synthesis and enhancement (arranged in ascending order of pixel values from the original bright image).

is able to govern them into a curve. Of particular interest is the results from enhancing synthesized low-light image in Fig. 5.9, where the distribution of the $\mathcal{GP}$ enhancement (the red scattered distribution) closely matches that of the original image's pixel values (the blue curve). This, including the obvious visual superiority proves the effectiveness and importance of localized enhancement brought upon by the $\mathcal{GP}$.

Next, the contrast enhancement results of the proposed method ($\mathcal{GP}$(prior1)) were evaluated in comparison to LACE, FBE, WVM, and LIME. The proposal's results in Fig. 5.10 show better quality in terms of contrast and sharpness with minimum noise on real low-light image. In the regions bounded by the blue boxes, it can seen that all methods result in better contrast after the enhancement. However, in the output of the CNN only model, many of the details were lost and the image appears blurred. The proposed method

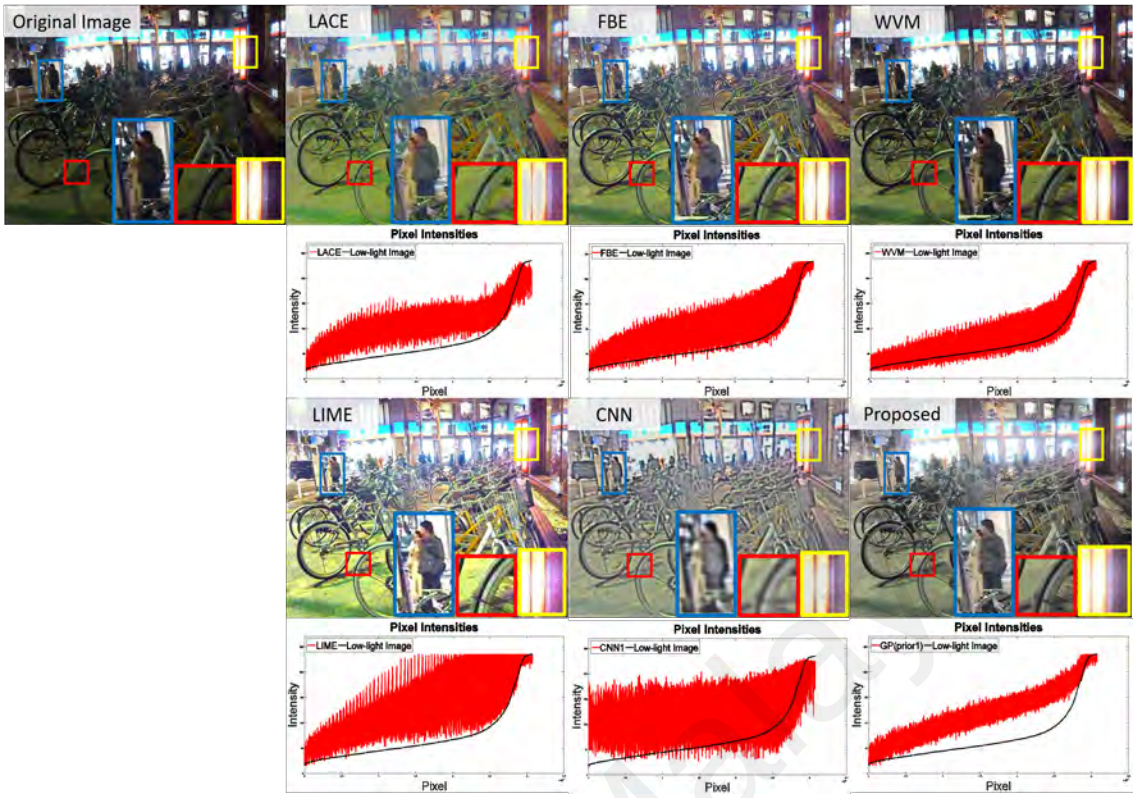


Figure 5.10: Example of the contrast enhancement on a real low-light image, and the intensity of each pixel before and after enhancement of the respective methods (arranged in ascending order of pixel values from the original low-light image).

and LIME show the best contrast and sharpness, for example, the bicycle wheel bounded by the red boxes. However, LIME's result has more noise and color distortion as shown in the yellow bounding boxes. It can be observed that the area next to the light source appears noisy with unnatural pinkish hue whereas the proposed method's results look more natural.

This observation is similar to the results of the synthesized low-light image in Fig. 5.11. The cow in the red bounding box and edges of the tree in the yellow bounding box of the $\mathcal{GP}$'s result have better contrast than FBE and WVM whereas both LACE and LIME suffer from color distortions, such as the clouds in the blue bounding box where both show a purplish hue. While both LACE and LIME have aesthetically more pleasing enhancements contributed by their vibrant colors, the details retrieved by the proposed method are still comparable even though not as apparent because human perceptions are sensitive to colors. Thus, in attempt to further justify that the proposal indeed enhances the darkened

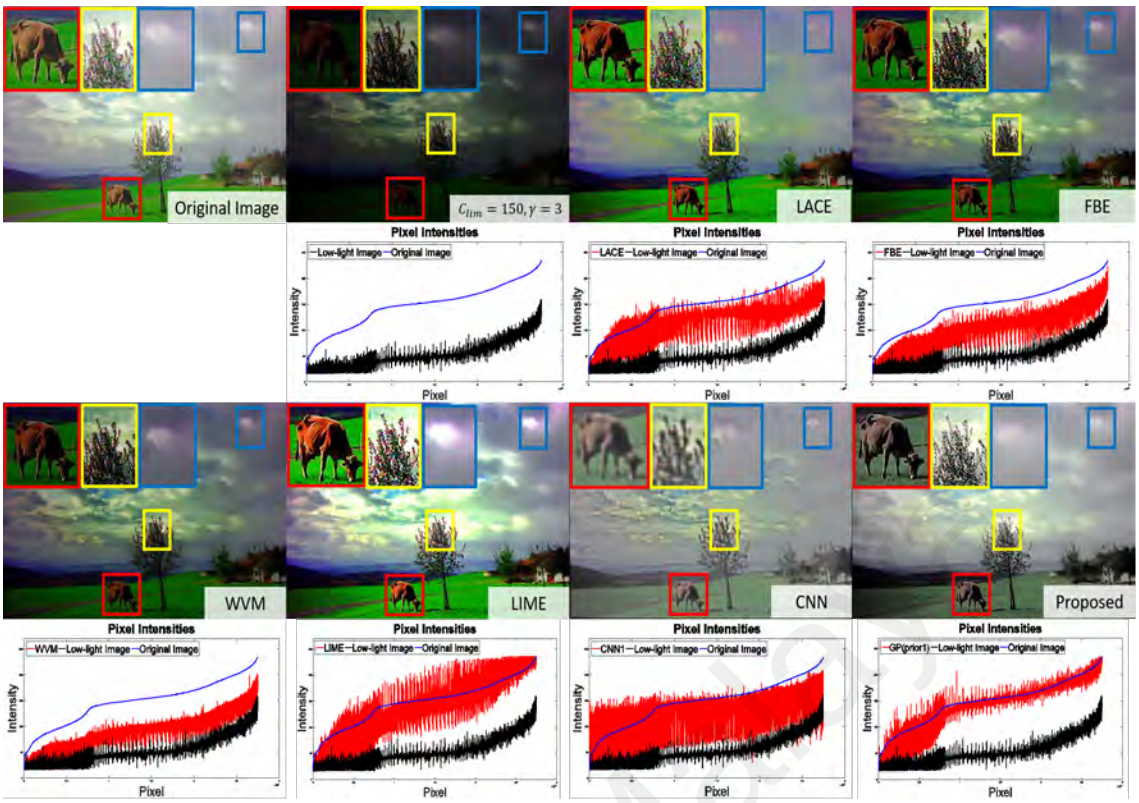


Figure 5.11: Example of the contrast enhancement on a synthesized low-light image, and the intensity of each pixel before and after synthesis and enhancement (arranged in ascending order of pixel values from the original bright image).

sample towards the original image, the pixel intensity distributions are shown in Fig. 5.10 and 5.11. The $\mathcal{GP}$ results' are very much less scattered than LACE and CNN, while the distributions of FBE and WVM are very much lower than the pixel value of the original bright image (blue curve) indicating under enhancement. As for LIME, the distribution is able to match the original bright image but is relatively more scattered, moreover, in the experiments, it is found to fail in certain images as shown in Fig. 5.13. Especially notable for the synthesized image in Fig. 5.11, the proposed $\mathcal{GP}$ model produces a distribution that closely matches the original bright image. Additional examples of results are shown in Fig. 5.12-5.13 for both real and synthetic low-light images respectively, and more results are found in Appendix D. It should be noted that even though the proposed enhancement results may lack in color vibrancy as compared to LIME, it is not a hindrance to the target for higher level applications in the quantitative evaluations.

Figure 5.12: Contrast enhancement results of real low-light images.

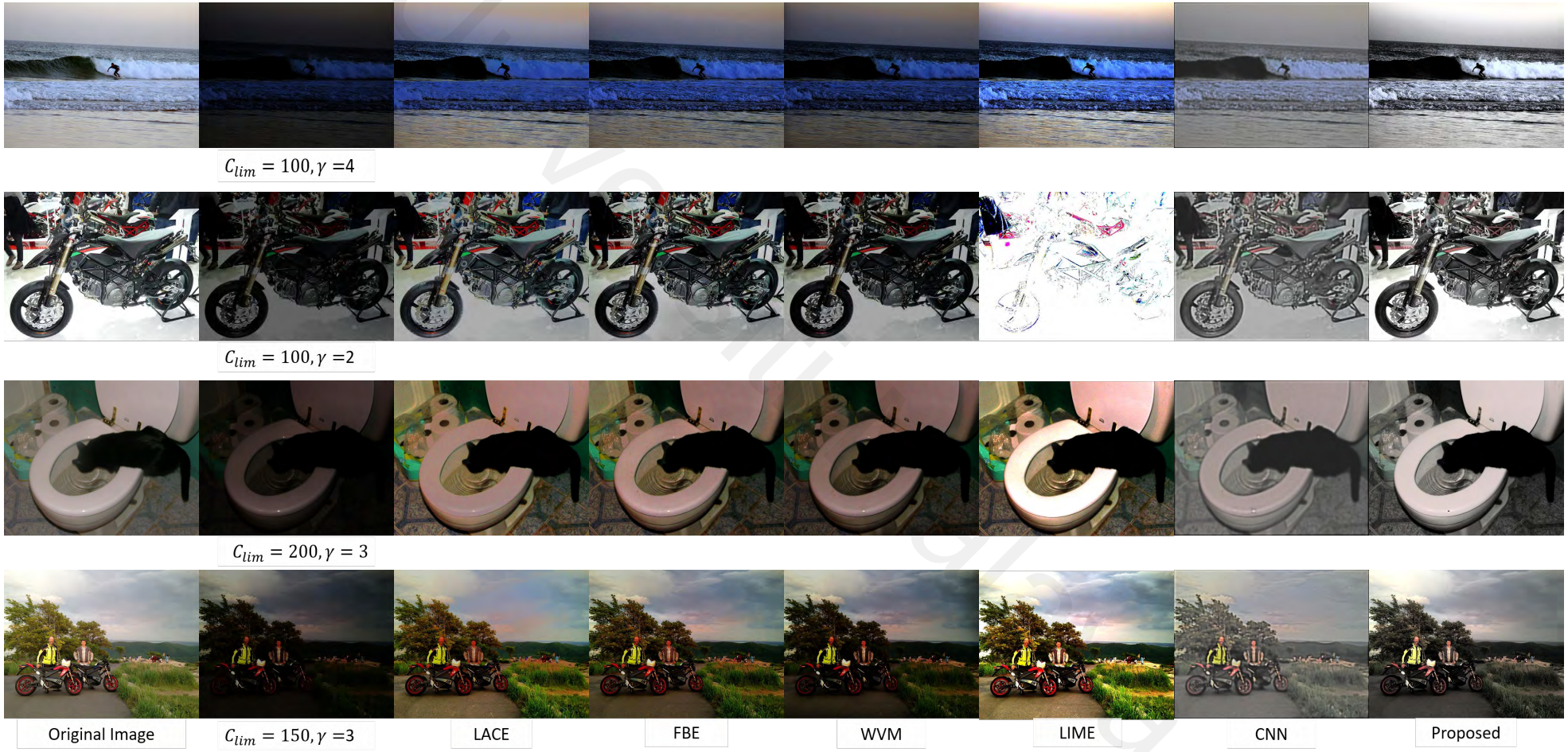


Figure 5.13: Contrast enhancement results of synthesized low-light images.

Table 5.1: Average PSNR results.

Approach	Darkened	LACE	FBE	WVM	LIME	CNN1	$\mathcal{GP}$(prior1)	CNN2	$\mathcal{GP}$(prior2)
PSNR (dB)	10.44	14.95	14.68	12.88	15.08	15.88	16.25	**16.42**	16.10

Table 5.2: Computational time.

Approach	LACE	FBE	WVM	LIME	CNN1	$\mathcal{GP}$(prior1)	CNN2	$\mathcal{GP}$(prior2)
Time (s)	15.97	0.17	3.03	**0.07**	0.13	1.25	0.13	1.29

5.3.4 Quantitative Evaluation

The quantitative assessments are carried out on 3 evaluation metrics: the PSNR, local features matching, and l_1-norm luminance histogram distance. The measures shown are from the synthetically darkened test images, the approaches by LACE, FBE, WVM, and LIME, the CNN, and the proposed $\mathcal{GP}$ framework.

5.3.4 (a) PSNR

In simple terms, the PSNR calculates how well the pixels of an image matches the pixels of the reference image. Given the enhanced low-light image I and the bright image reference as R, the metric is calculated using the MSE (Eqn. 5.16) as follows:

$$\text{PSNR} = 10\log_{10}\left(\frac{\max(R)^2}{\mathcal{L}(R,I)}\right) \tag{5.19}$$

The unit of measurement for the PSNR is the logarithmic scale decibel (dB) where the higher value indicates a better match, and thus a better enhancement result.

Table 5.1 shows the average PSNR results calculated for all RGB channels of the tested images. Both the CNN models and the proposed method outperform the state-of-the-art solutions with satisfactory computation time, as shown in Table 5.2. However, it is obviously in conflict with the qualitative assessment in Fig. 5.10 and 5.11, where LACE, FBE, and WVM produced images with better visual quality than the CNN models. Figure 5.14 additionally shows examples enhanced by all the methods where the PSNR displays

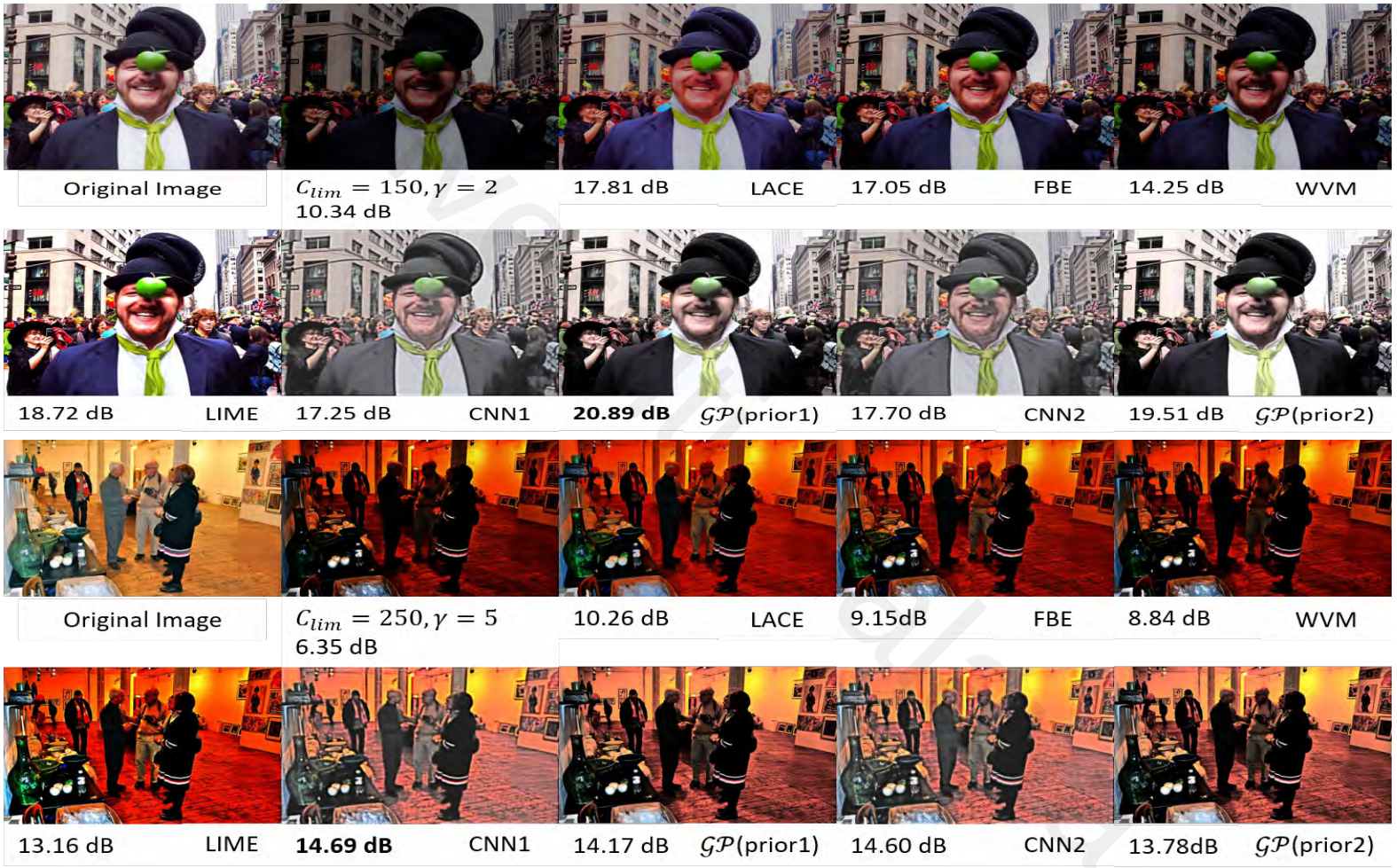


Figure 5.14: PSNR of synthetic low-light images enhancements.

inconsistency in evaluating the quality. While the PSNR for the CNN results are relatively low in the top example of Fig. 5.14, both CNN1 and CNN2 record the highest PSNR for second example in Fig. 5.14 respectively even though the results are clearly unnatural.

This is significant indicator that PSNR is not the ideal measure for low-light image enhancement. Moreover, such quality measurements do not bring out the significance of low-light image contrast enhancement for higher level application such as detection and tracking. Therefore, new metrics are introduced in this work, namely the local feature matching and the histogram l_1-norm distance, to evaluate the ability of enhancement algorithms to improve valuable details of low-light images.

5.3.4 (b) Local Features Matching

Before the breakthrough of learned features, local features were the forerunners for detection and recognition tasks, and are still used in part to this day (Khan et al. (2012); Mottaghi et al. (2015); S. Zhang et al. (2015)). Hence, detected local features are capitalized as a gage for useful information content retrieved by enhancing low-light images. Furthermore, the reliability of this measure is heightened by matching features detected from the enhancement results to the original bright image to ensure the retrieved details are not “false" features from noise and artifacts created by the enhancements. The precision Pr, recall Rc, and F-score were then calculated based on information retrieval context as follows:

$$\begin{aligned} Pr &= \frac{|q_{rlv} \cap q_{rtv}|}{q_{rtv}}; \\ Rc &= \frac{|q_{rlv} \cap q_{rtv}|}{q_{rlv}}; \\ F_\beta\text{-score} &= (1+\beta^2)\frac{Pr \cdot Rc}{(\beta^2 \cdot Pr) + Rc}, \end{aligned} \tag{5.20}$$

Table 5.3: Average precision, recall, and F-scores of feature matching

Approach	Precision	Recall	F_1-score	F_2-score
Darkened	0.4514	0.1711	0.2090	0.1824
LACE	**0.6358**	0.4305	0.4820	0.4445
FBE	0.5959	0.4659	0.4831	0.4639
WVM	0.5794	0.3458	0.3722	0.3496
LIME	0.3205	0.6463	0.4076	0.5062
CNN1	0.3168	0.3779	0.2872	0.3221
$\mathcal{GP}$(prior1)	0.4745	**0.6563**	**0.5292**	**0.5871**
CNN2	0.3815	0.4311	0.3699	0.3952
$\mathcal{GP}$(prior2)	0.4664	0.6474	0.5202	0.5818

where q_{rlv} refers to feature points extracted from the original bright image and q_{rtv} are feature points from the enhanced image, while $|q_{rlv} \cap q_{rtv}|$ indicates the correctly matched points. β is the weight variable for the precision and recall in computing the F-score. A higher β puts more weight on recall than precision.

The Scale Invariant Feature Transform (SIFT) (Lowe (2004)) is used as the local feature for this evaluation. SIFT is a scale, rotation, illumination, and viewpoint invariant feature that extract interest points from an image. Separated into two components, the detector and descriptor, the detector implements the difference of Gaussian to locate key-points based on edge response, whereas the descriptor describes the detected keypoints by calculating the gradient magnitude and orientation to form orientation histograms. The result is a 128 dimension vector that is robust to many transformations and descriptive for applications like object recognition and image matching. This reason, coupled with its lightweight computational requirement makes it a suitable feature for this evaluation metric.

In the experiments, the peak threshold parameter for SIFT was set to be 10 so as to only remain points detected from regions with strong contrast, and correctly matched points are defined as features with matching descriptors, location, scale and orientation. In order to emphasize on the useful information that enhancement can retrieve, the recall is more weighted with $\beta = 2$ (F_2-score). Table 5.3 shows that the proposed method out-

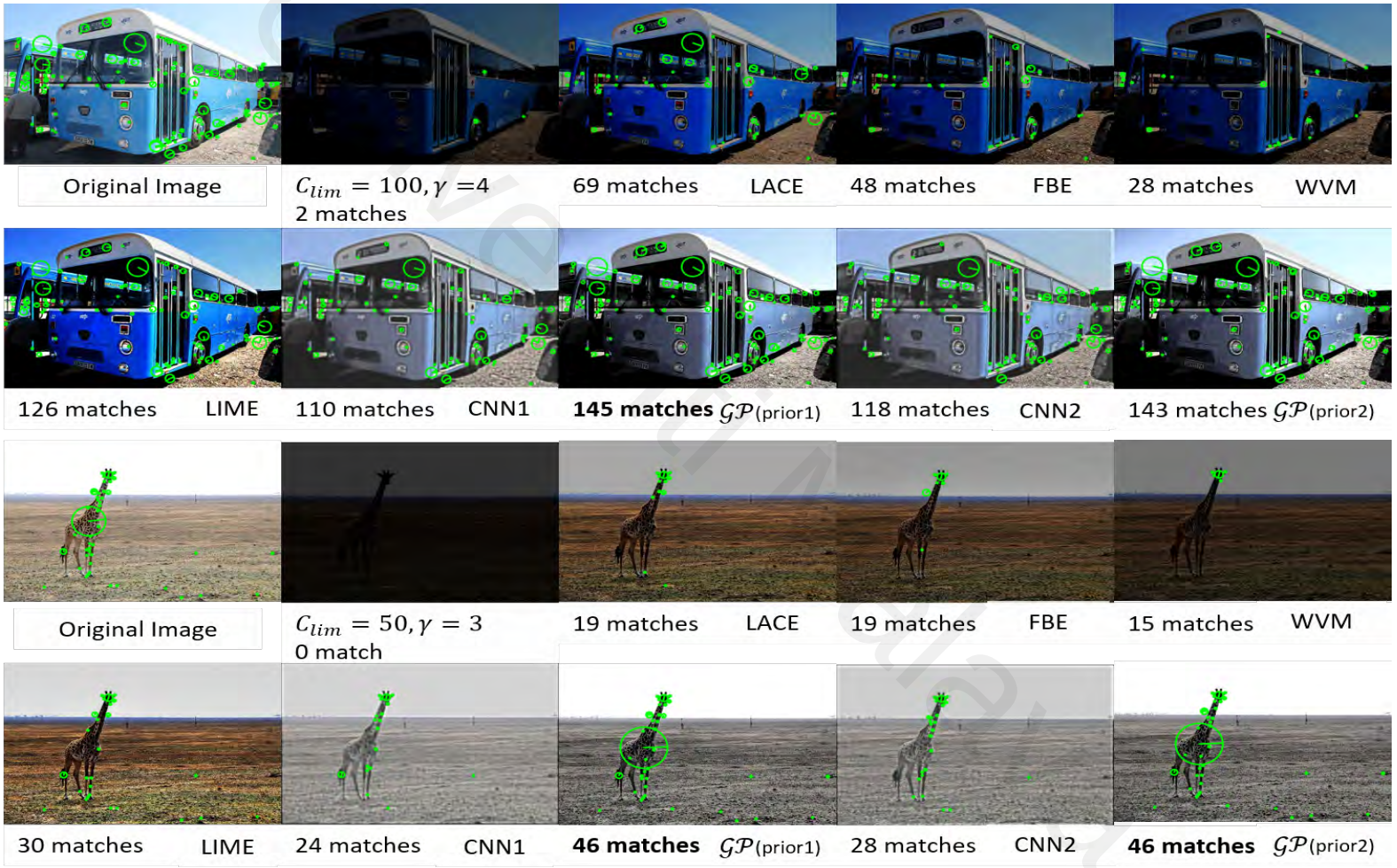


Figure 5.15: SIFT features matched in synthetic low-light images using different enhancements methods.

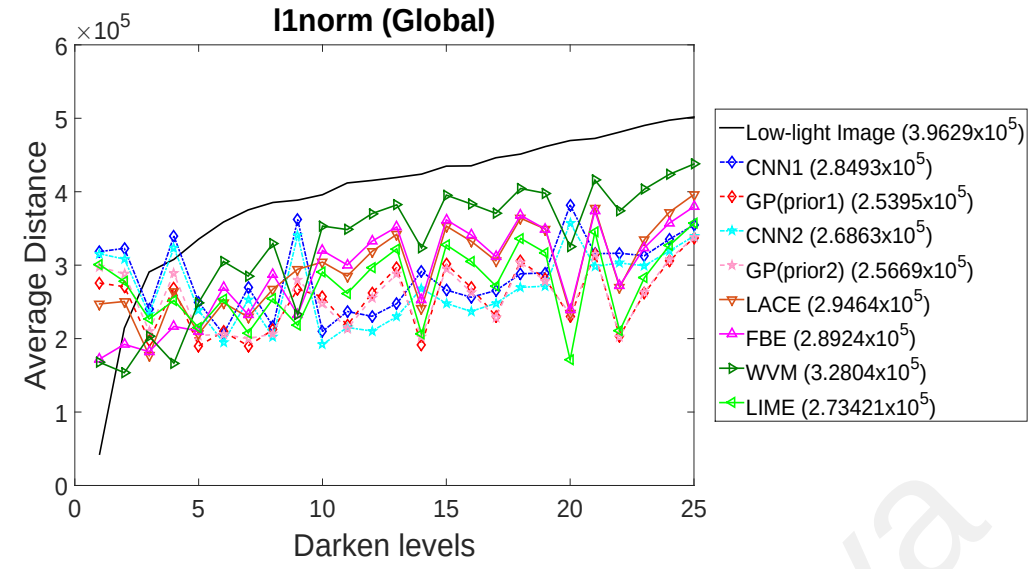
performs all methods in recall and F_2-score. It was analyzed and found that the LACE method has the best precision because it retrieves less local features in total. Even so, for the average scores where at equal weights for precision and recall in the F_1-score, the proposed method is still the highest scoring.

Figure 5.15 shows some examples of the features detected from the enhancement results and matched to the features detected in the original bright images. The darkening of the images significantly impacts the features extractable from the objects, particularly in the second example of Fig. 5.15 where the synthesized low-light image does not have any features matched. Nevertheless, each enhancement method is able to retrieve some features, particularly, the proposed $\mathcal{GP}$ with the most retrievals. These examples and the scores effectively show that the proposed approach retrieve more relevant local features for further use.

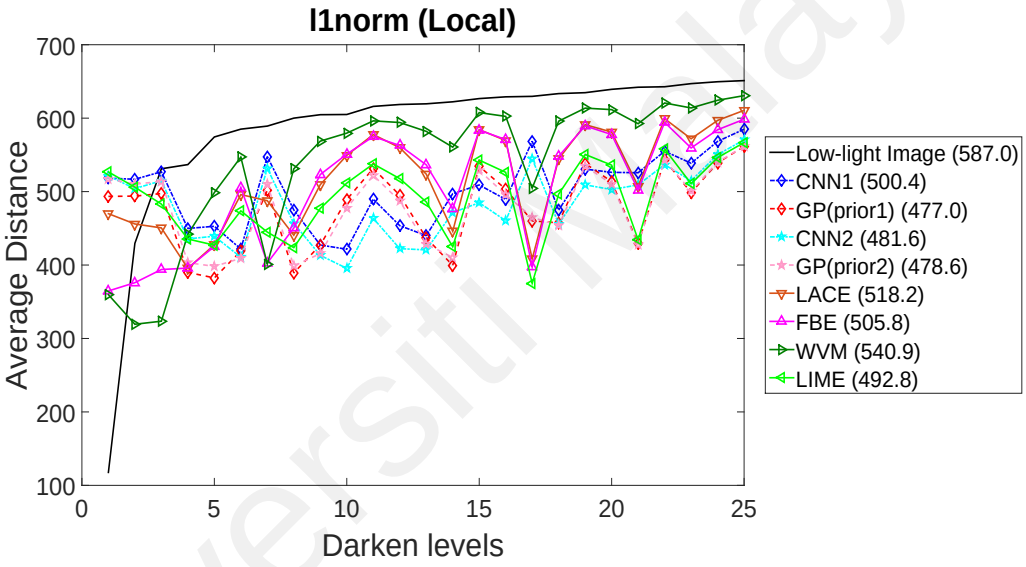
5.3.4 (c) l_1-norm

The second evaluation method proposed is the histogram distance measure using l_1-norm. Color histogram matching is commonly employed in tracking applications (Smeulders et al. (2014)). Therefore, this assessment serves to evaluate the prospect of enhancement results for tracking algorithms based on histogram similarity. The l_1-norm is calculated between color histograms of the original bright image and the enhanced images. The comparison was performed on both global image intensity histogram and the histograms of local patches in the image. The local histograms were obtained by dividing the image into non-overlapping patches of 32×32 pixels and then calculate the average distance of all the patches. The intensity histograms were set to have 32 bins for the assessment.

Figure 5.16 shows the comparison results divided into 25 darken levels, referring to the combinations of the C_{lim} and γ parameters in generating the synthetic low-light images, and arranged in ascending order of distances between the synthetic low-light



(a)



(b)

Figure 5.16: Comparison of l_1-norm of intensity histograms with 32 bins for (a) global image intensities (b) local 32×32 pixels patch intensities. Values (in the brackets) of the legends indicate total average distances.

images and their respective original bright images. All of the methods are able to shorten the distances for both local and global color histograms, except for levels below 3, where the darkening is not severe. It is also for these initial levels where LACE, FBE, and WVM perform well, but they gradually decline as the level increases. The proposed method ($\mathcal{GP}$(prior1) and $\mathcal{GP}$(prior2)) consistently perform the best for levels above 5, except for two levels where LIME is better. Even though the proposed method does not explicitly enhance the color content like the others, it still outperforms them on average

for this measure. Not only does this show the potential of the proposal to support low-light tracking operations, but also suggests that the method brings the image closer to the original state, although less pleasing to human observation.

5.3.5 Public Datasets

This section shows the comparison between the state-of-the-arts with the proposed method on 4 public datasets related to lighting research that were discussed in Section 2.1.3, Phos (Vonikakis et al. (2013)), DALI (Simo-Serra et al. (2015)), Webcam (Verdie et al. (2015)), and ALCN-2D (Rad et al. (2017)). Figures 5.17 - 5.20 show example results from each of these datasets and it can be seen that the proposed method fare rather well against the others.

Inspecting the results in the Phos image of Fig. 5.17, the proposed method shows a good contrast enhancement, as seen in the enlarged regions, without over-enhancing bright regions (red box of first example). Moreover, the color of the results most closely resemble the baseline image among all the methods, but with even better contrast than the recommended baseline.

On the other hand, for the DALI dataset, the result of the $\mathcal{GP}$ shows a clear improvement of the contrast as shown in the areas bounded by the colored boxes of the images in Fig. 5.18. Though it is noted that the result darkens the shadowed areas in order to accentuate the details.

Figures 5.19 - 5.20 are examples from the Webcam and ALCN-2D datasets respectively. It can be seen that the proposed method is able to handle bright regions, without over-enhancing them like the LIME method. Additionally, these two examples show that the proposal do not amplify the noise content like the LACE method. Moreover, the state-of-the-art methods tend to show a pinkish hue in their results, particularly seen in the second example, bounded by the red boxes of Fig. 5.20.

Figure 5.17: Contrast enhancement results of images from Phos dataset.

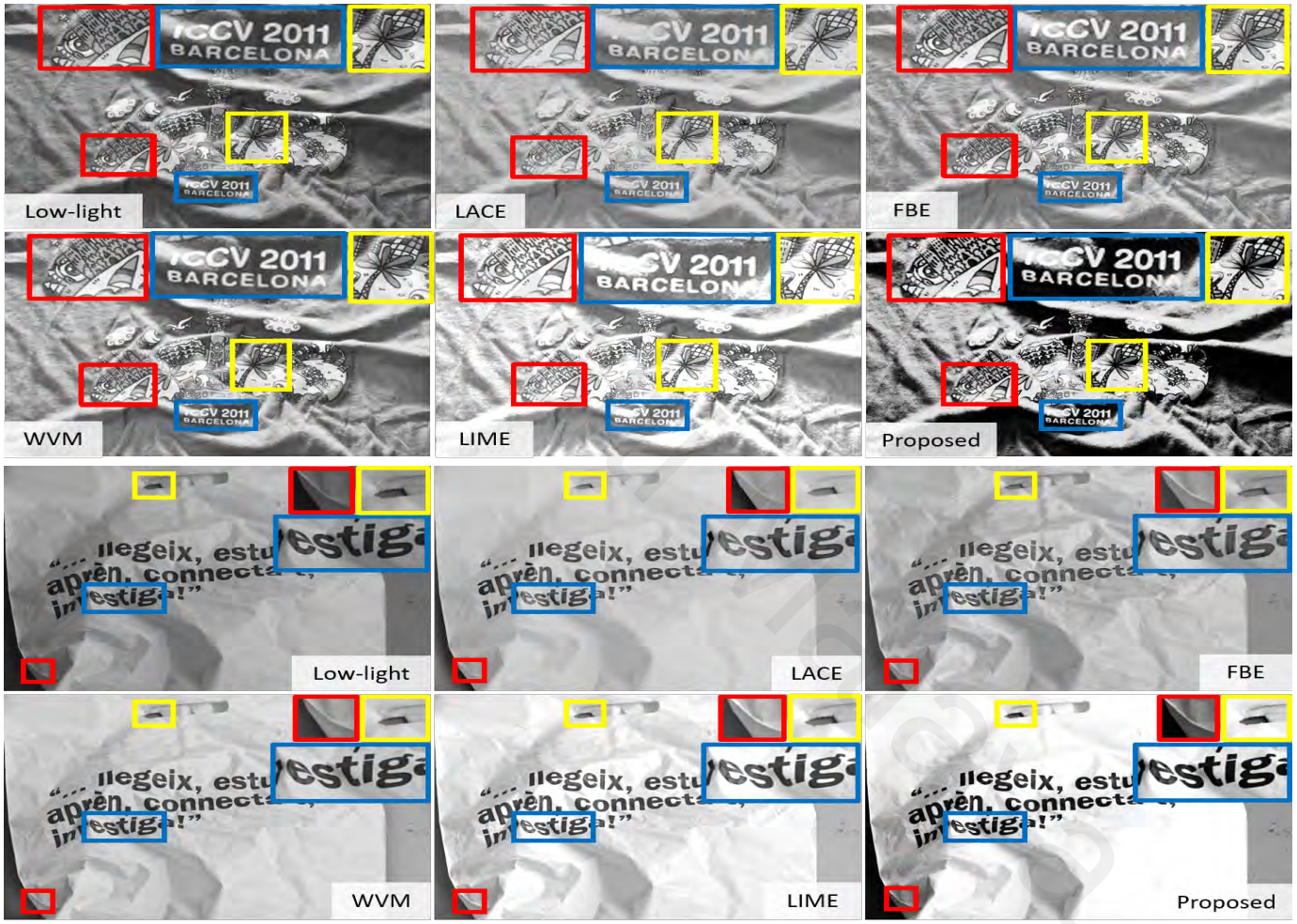


Figure 5.18: Contrast enhancement results of images from DALI dataset.

Figure 5.19: Contrast enhancement results of images from Webcam dataset.

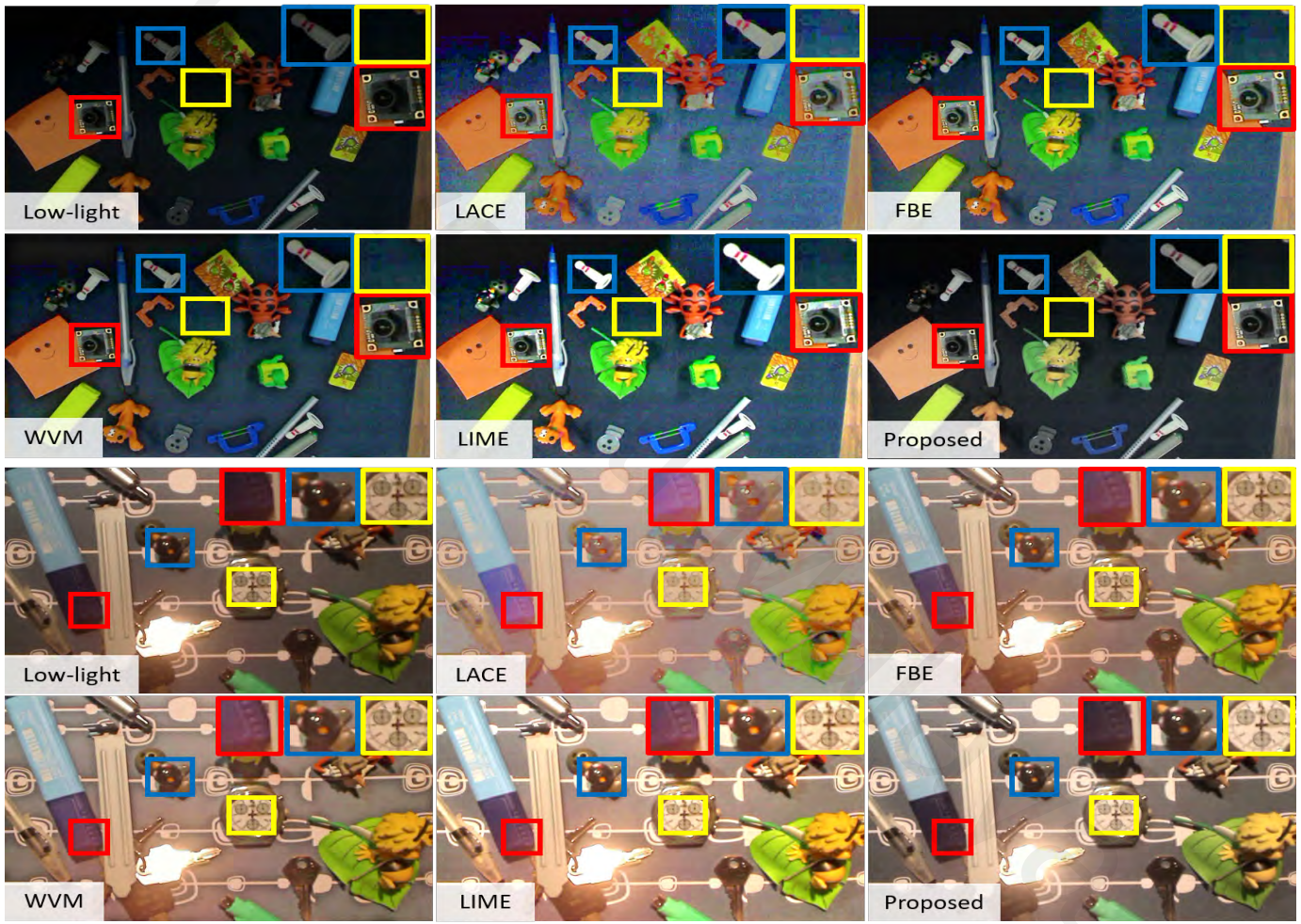


Figure 5.20: Contrast enhancement results of images from ALCN-2D dataset.

5.4 Summary

In summary, this chapter addresses the low-light challenge through contrast enhancement from the perspective of information retrieval instead of aesthetic restoration. Considering the challenge in both global and local luminance variations, the low-light image enhancement is modeled as a distribution of localized enhancement functions using the $\mathcal{GP}$ assisted by a CNN as an intermediate model. The CNN is trained using large data synthesized based on luminance statistics of real images so that it generates a globally optimized training data for the $\mathcal{GP}$.

Taking into account the objective of retrieving informative details that would contribute to higher level object-based applications, new information retrieval assessments are introduced to point out the practicability of the enhancement results. The proposed framework is a new approach that is unlike typical low-light image enhancement methods and it outperforms the state-of-the-art qualitatively and quantitatively in both the conventional and the new evaluation metrics.

CHAPTER 6: CONCLUSIONS

This chapter concludes the research work of this thesis with a summary of the work done and findings, the limitations of the current work, and possible directions for future progress.

6.1 Summary

This research work aims at looking into the low-light phenomenon in computer vision, with three distinct objectives. The first is to establish a standard benchmarking dataset for this work and also for the other research work in the low-light domain because of the glaring lack of a standard data to encourage research in this direction. To this end, the ExDark low-light object image dataset is proposed, containing 7,363 low-light images with a first of its kind categorization into 10 types of illumination conditions. This new categorization could lay the groundwork for image processing works in low-light that has never been done before, as it brings to light the many different characteristics that were previously assigned into a singular group of "low-light". In particular, low-light enhancement research, that has consistently faced the over or under enhancement problem due to irregular illumination, can benefit from this illumination classification in the effort to design more robust algorithms. Additionally, the dataset is annotated with 12 common object classes including bounding box level annotation which also makes it the first object image dataset that consist fully of low-light images. This will enable and encourage the growth of research and development of applications such as object detection and recognition in low-light. Thus, this contribution achieves the first objective by addressing the lack of data problem and promote progress of low-light research.

The second objective is to gain insights on the characteristics of low-light to guide the progression of computer vision in this direction. Through low-level analysis, that is by

studying the global intensity histograms and local region intensities, it is found that low-light image shows distinct illumination variations that are not only different from bright images, but also diverse in terms of low-light image types. Using the new ExDark dataset with the illumination categories, the global intensity histograms of the images show subtle differences in their distributions based on their types. Whereas, local region intensities showed that the illumination levels are locally irregular within a low-light image due to the presence of light source(s) in certain low-light environment, a characteristic that is found in the *single, weak, strong, window,* and *screen* types. These findings could be especially useful for the low-light enhancement works where the understanding of such traits would aid in designing robust algorithms that address such irregularities for effective enhancement. On the other hand, high-level analysis using hand-crafted and learned features had shown that the current state-of-the-arts of object based features are insufficient to solve the challenges presented by low-light conditions. In the analysis using prominent hand-crafted features, edges, gradients, and superpixels, for object localization, they noticeably show difficulty to provide accurate localizations in low-light images. This is because the low illumination and contrast cause objects to appear to blend to the background which subsequently cause the lack of such features. Noting such shortcoming from this analysis, future efforts to design feature extractors or object detectors would have to take this finding into consideration. As for features learned using the state-of-the-art of deep learning models, the Resnet-50, in object classification, the analyses showed that the features of an object is altered by low-light. This is counter-intuitive to the common notion that learned features disregard illumination effects as they capture higher level abstract features, but as seen from the analysis, same object class gives different features in bright and low-light conditions. Furthermore, a study on the attention maps of the learned model showed that the irregular lighting of low-light environment distracts the model and causes misclassification, indicating that current learning models are still lacking. These findings serve as

the second contribution of this thesis to not only show the challenges brought by low-light as intended by the second objective, but also show the many directions where low-light research can be expanded.

Lastly, the third objective of this thesis is to develop low-light image enhancement framework that primarily retrieves features while maintaining fair visual quality, a target that is different from the conventional aims of enhancement works. This objective was approached based on the insights gained from the aforementioned analyses, particularly the understanding of global and local illumination variations present in low-light image. Hence, a low-light image contrast enhancement framework, using $\mathcal{GP}$ is proposed to retrieve useful features that would assist object-based computer vision tasks. This is in consideration of the localized illumination variation of low-light images as well as the need to emphasize object features, therefore specific functions for each local pixels or regions is necessary to reach optimal enhancement. This is achievable by the $\mathcal{GP}$ as it models data into a distribution of functions, i.e. localized enhancement functions are governed within a single distribution. Subsequently, two new evaluation metrics, the local features matching and l_1-norm distance measure of intensity histogram, are proposed as well to benchmark the practicability of contrast enhancement algorithms in supporting detection and tracking applications respectively. The proposed method outperforms the state-of-the-art low-light enhancement algorithms in the conventional PSNR measure by 1.17dB, as well as the new evaluations where the F_1-score of features matching is improved by 9.5% and the l_1-norm distance of global and local intensity histograms are reduced by 7.1% and 3.2% respectively. These results asserts the proposed method as an algorithm that not only performs visual enhancement on low-light images, but also retrieves features that could support applications, thus achieving the third objective.

6.2 Limitations

This research work possesses a few limitations. Firstly, while the proposed dataset, the ExDark is the first of its kind with comparatively large amount of low-light data to any object datasets available now, it is still far from the sheer numbers of bright data provided by the current popular object datasets. This serves as barrier for works that require big data, such as deep learning approaches, where currently there is a reliance on pre-training on other larger data in order to work on smaller datasets such as this. As discussed in Section 1.1, the significant breakthrough of deep learning for object classification was by using the ImageNet data which at the time already amassed a staggering total of over 1 million images. Moreover, visual data is a representation of the visual environment that people live in, hence, a larger amount of data is invaluable to demonstrate and provide a reliable representation of real world environments, and subsequently guide computer vision research to reach human vision capability and beyond.

The next limitation is in the proposed method where the computational time and color content is still somewhat lacking. The current $\mathcal{GP}$ model is computationally intensive and the time to process a single image takes more than 1 second as shown by Table 5.2 in Section 5.3.4 (a). This is because the $\mathcal{GP}$ is significantly more complex than other statistical or transformation models. Unlike statistical models that bin pixels into histograms to be easily modified, or transformation models that has a ready model and set parameters to process the pixels, the $\mathcal{GP}$ builds exclusive distributions for each image it enhances. Furthermore, as discussed in Section 5.2.1, the $\mathcal{GP}$ computes the distribution of functions based on covariance functions, a dissimilar yet more computationally demanding approach than general curve fitting regression models. Hence, the current model of $\mathcal{GP}$ for low-light image enhancement still leaves room for optimization before it can reach real-time performance where low-light systems such as surveillance truly

matters.

In regards to the color content, the current model does not improve this aspect as it is of less importance to the targeted task. As stated in Section 5.3.1, the proposed model uses the YC_bC_r color space and only enhances the luminance (Y) channel while the chrominance components (C_bC_r) remain untouched. This choice of action is also due to the computation requirement that has made inclusion of the color into the enhancement impractical. Nonetheless, this has caused rapid deterioration of the color especially in very low-light conditions, where not only do the color reduce due to lack of light transmission, it is reliant on the conversion from the RGB space to the YC_bC_r space, and the decomposition into the respective channels. Therefore, the enhancement outcomes of the proposed model appears less pleasing to the human observation and could affect more fine-grained applications such as recognition tasks that require identification by color.

6.3 Future Works

Taking into consideration the limitations of the current stage, there are various directions that can be taken to progress this line of research forward. One of which is to expand the dataset of low-light object images. As discussed in Section 6.2, the amount of low-light data is still very much smaller than bright object image datasets. Expanding the dataset would be able to encourage more research efforts and overcome the lack of data issue that plagues the research community especially those that work on or intend to implement deep learning for low-light research.

On the other hand, a potential direction is in the development and improvement of the solution for low-light object applications. One of which is to optimized the currently proposed model for real time applications. Alternatively, deep neural network approaches hold great potential as well in solving low-light challenges considering the successes shown in other enhancement works (C. Dong et al. (2015); Jain & Seung (2009); Lars-

son et al. (2016)). Moreover, Generative Adversarial Network (GAN) by Goodfellow et al. (2014) has also been gaining traction in the computer vision community and could potentially be valuable for the low-light domain considering its ability in estimating the distribution of input data. This trait could be useful for low-light image enhancement as it might be possible for a GAN to learn individualistic distributions for different low-light conditions. Moreover, the adversarial training scheme that is implemented to optimize the model could very well reduce the uncertainties brought upon by the diminished vision capability of humans in low-light conditions.

REFERENCES

Anarkooli, A. J., & Hosseinlou, M. H. (2016). Analysis of the injury severity of crashes by considering different lighting conditions on two-lane rural roads. *Journal of Safety Research*, *56*, 57–65.

Bilodeau, G.-A., Torabi, A., St-Charles, P.-L., & Riahi, D. (2014). Thermal–visible registration of human silhouettes: A similarity measure performance evaluation. *Infrared Physics & Technology*, *64*, 79–86.

Calabrese, C., Mejia, B., McInnis, C. A., France, M., Nadler, E., & Raslear, T. G. (2017). Time of day effects on railroad roadway worker injury risk. *Journal of Safety Research*, *61*, 53–64.

Cheng, M.-M., Zhang, Z., Lin, W.-Y., & Torr, P. (2014). Bing: Binarized normed gradients for objectness estimation at 300fps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3286–3293).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 1, pp. 886–893).

Davis, J. W., & Keck, M. A. (2005). A two-stage template approach to person detection in thermal imagery. In *2005 IEEE Workshops on Application of Computer Vision (WACV/MOTIONS)* (Vol. 1, pp. 364–369).

Davis, J. W., & Sharma, V. (2007). Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, *106*(2), 162–182.

de Melo, S. N., Pereira, D. V., Andresen, M. A., & Matias, L. F. (2017). Spatial/temporal variations of crime: a routine activity theory perspective. *International Journal of Offender Therapy and Comparative Criminology*, 0306624X17703654.

Dollár, P. (n.d.). *Piotr's Computer Vision Matlab Toolbox (PMT)*. `https://github.com/pdollar/toolbox`.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning (ICML)* (pp. 647–655).

Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep

convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 295-307.

Dong, J., Ge, J., & Luo, Y. (2007). Nighttime pedestrian detection with near infrared using cascaded classifiers. In *2007 IEEE International Conference on Image Processing (ICIP)* (Vol. 6, pp. VI–185).

Dong, X., Wang, G., Pang, Y., Li, W., Wen, J., Meng, W., & Lu, Y. (2011). Fast efficient algorithm for enhancement of low lighting video. In *2011 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6).

Elguebaly, T., & Bouguila, N. (2013). Finite asymmetric generalized gaussian mixture models learning for infrared object detection. *Computer Vision and Image Understanding*, *117*(12), 1659–1671.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, *111*(1), 98–136.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010, June). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Fang, Z., Cao, Z., Xiao, Y., Zhu, L., & Yuan, J. (2016). Adobe boxes: Locating object proposals using object adobes. *IEEE Transactions on Image Processing*, *25*(9), 4116–4128.

Fei-Fei, L., Fergus, R., & Perona, P. (2007). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, *106*(1), 59–70.

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).

Fotios, S., Unwin, J., & Farrall, S. (2015). Road lighting and pedestrian reassurance after dark: A review. *Lighting Research & Technology*, *47*(4), 449–469.

Fu, H., Ma, H., & Wu, S. (2012). Night removal by color estimation and sparse representation. In *2012 IEEE International Conference on Pattern Recognition (ICPR)* (pp. 3656–3659).

Fu, X., Zeng, D., Huang, Y., Liao, Y., Ding, X., & Paisley, J. (2016). A fusion-based enhancing method for weakly illuminated images. *Signal Processing*, *129*, 82–96.

Fu, X., Zeng, D., Huang, Y., Zhang, X.-P., & Ding, X. (2016). A weighted variational model for simultaneous reflectance and illumination estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2782–2790).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 2672–2680).

Griffin, G., Holub, A., & Perona, P. (2007). Caltech-256 object category dataset.

Guo, X., Li, Y., & Ling, H. (2017). Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, *26*(2), 982–993.

Hanaoka, K. (2016). New insights on relationships between street crimes and ambient population: Use of hourly population data estimated from mobile phone users' locations. *Environment and Planning B: Planning and Design*, 0265813516672454.

He, K., Sun, J., & Tang, X. (2011). Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(12), 2341–2353.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

Huang, S.-C., Cheng, F.-C., & Chiu, Y.-S. (2013). Efficient contrast enhancement using adaptive gamma correction with weighting distribution. *IEEE Transactions on Image Processing*, *22*(3), 1032–1041.

Jain, V., & Seung, S. (2009). Natural image denoising with convolutional networks. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 769–776).

Kang, D., Han, H., Jain, A. K., & Lee, S.-W. (2014). Nighttime face recognition at large standoff: Cross-distance and cross-spectral matching. *Pattern Recognition*, *47*(12), 3750–3766.

Kaur, M., Kaur, J., & Kaur, J. (2011). Survey of contrast enhancement techniques based on histogram equalization. *International Journal of Advanced Computer Science and Applications*, *2*(7), 137–141.

Khalilikhah, M., & Heaslip, K. (2017). Improvement of the performance of animal crossing warning signs. *Journal of Safety Research*, *62*, 1–12.

Khan, F. S., Anwer, R. M., Van De Weijer, J., Bagdanov, A. D., Vanrell, M., & Lopez, A. M. (2012). Color attributes for object detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3306–3313).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 1097–1105).

Land, E. H., et al. (1977). *The retinex theory of color vision*. Citeseer.

Larsson, G., Maire, M., & Shakhnarovich, G. (2016). Learning representations for automatic colorization. *arXiv preprint arXiv:1603.06668*.

Le Callet, P., & Autrusseau, F. (2005). *Subjective quality assessment irccyn/ivc database.* (http://www.irccyn.ec-nantes.fr/ivcdb/)

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, S. H., Chan, C. S., Mayo, S. J., & Remagnino, P. (2017). How deep learning extracts and learns leaf features for plant classification. *Pattern Recognition*, *71*, 1–13.

Leo, M., Medioni, G., Trivedi, M., Kanade, T., & Farinella, G. M. (2017). Computer vision for assistive technologies. *Computer Vision and Image Understanding*, *154*, 1–15.

Li, L., Wang, R., Wang, W., & Gao, W. (2015). A low-light image enhancement method for both denoising and contrast enlarging. In *2015 IEEE International Conference on Image Processing (ICIP)* (pp. 3730–3734).

Li, S. Z., Chu, R., Liao, S., & Zhang, L. (2007). Illumination invariant face recognition using near-infrared images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(4), 627–639.

Lim, H., Kim, C., Eck, J. E., & Kim, J. (2016). The crime-reduction effects of open-street cctv in south korea. *Security Journal*, *29*(2), 241–255.

Lim, J., Kim, J.-H., Sim, J.-Y., & Kim, C.-S. (2015). Robust contrast enhancement of noisy low-light images: Denoising-enhancement-completion. In *2015 IEEE International Conference on Image Processing (ICIP)* (pp. 4131–4135).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)* (pp. 740–755).

Loh, Y. P., & Chan, C. S. (2015). Unveiling contrast in darkness. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)* (pp. 266–270).

Lore, K. G., Akintayo, A., & Sarkar, S. (2017). Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, *61*, 650–662.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Łoza, A., Bull, D. R., Hill, P. R., & Achim, A. M. (2013). Automatic contrast enhancement of low-light images based on local statistics of wavelet coefficients. *Digital Signal Processing*, *23*(6), 1856–1866.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

Mahendran, A., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5188–5196).

Montoya, L., Junger, M., & Ongena, Y. (2016). The relation between residential property and its surroundings and day-and night-time residential burglary. *Environment and Behavior*, *48*(4), 515–549.

Mottaghi, R., Xiang, Y., & Savarese, S. (2015). A coarse-to-fine model for 3d pose estimation and sub-category recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 418–426).

Nayar, S., Nene, S., & Murase, H. (1996). Columbia object image library (coil 100). *Department of Comp. Science, Columbia University, Tech. Rep. CUCS-006-96*.

Nene, S. A., Nayar, S. K., Murase, H., et al. (1996). Columbia object image library (coil-20).

Olmeda, D., Premebida, C., Nunes, U., Armingol, J. M., & Escalera, A. d. l. (2013). Lsi far infrared pedestrian dataset.

Pedersen, E., & Johansson, M. (2016). Dynamic pedestrian lighting: Effects on walking speed, legibility and environmental perception. *Lighting Research & Technology*, 1477153516684544.

Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2008). Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).

Pour-Rouholamin, M., & Zhou, H. (2016). Investigating the risk factors associated with pedestrian injury severity in illinois. *Journal of Safety Research*, *57*, 9–17.

Qi, B., John, V., Liu, Z., & Mita, S. (2014). Use of sparse representation for pedestrian detection in thermal images. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)* (pp. 274–280).

Rad, M., Roth, P. M., & Lepetit, V. (2017). Alcn: Meta-learning for contrast normalization applied to robust 3d pose estimation. *arXiv preprint arXiv:1708.09633*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 91–99).

Rose, A. (1948). The sensitivity performance of the human eye on an absolute scale*. *Journal of the Optical Society of America*, *38*(2), 196–208.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... others (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*(1), 157–173.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Simo-Serra, E., Torras, C., & Moreno-Noguer, F. (2015). Dali: deformation and light invariant descriptor. *International Journal of Computer Vision*, *115*(2), 136–154.

Smeulders, A. W., Chu, D. M., Cucchiara, R., Calderara, S., Dehghan, A., & Shah, M. (2014). Visual tracking: An experimental survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1442–1468.

Tompson, L., & Bowers, K. (2013). A stab in the dark? a research note on temporal patterns of street robbery. *Journal of Research in Crime and Delinquency*, *50*(4), 616–631.

Tong, S., Loh, Y. P., Liang, X., & Kumada, T. (2016). Visual attention inspired distant view and close-up view classification. In *2016 IEEE International Conference on Image Processing (ICIP)* (pp. 2787–2791).

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1521–1528).

Torralba, A., Fergus, R., & Freeman, W. T. (2008). 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, *30*(11), 1958–1970.

Van Der Maaten, L. (2014). Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, *15*(1), 3221–3245.

Vangorp, P., Myszkowski, K., Graf, E. W., & Mantiuk, R. K. (2015). A model of local adaptation. *ACM Transactions on Graphics*, *34*(6), 166:1–13.

Verdie, Y., Yi, K., Fua, P., & Lepetit, V. (2015). Tilde: a temporally invariant learned detector. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5279–5288).

Vonikakis, V., Chrysostomou, D., Kouskouridas, R., & Gasteratos, A. (2013). A biologically inspired scale-space for illumination invariant feature detection. *Measurement Science and Technology*, *24*(7), 074024.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., & Gong, Y. (2010). Locality-constrained linear coding for image classification. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3360–3367).

Wang, S., Zheng, J., Hu, H.-M., & Li, B. (2013). Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, *22*(9), 3538–3548.

Williams, C. K., & Rasmussen, C. E. (2006). Gaussian processes for machine learning. *the MIT Press*, *2*(3), 4.

Winn, J., Criminisi, A., & Minka, T. (2005). Object categorization by learned universal visual dictionary. In *Computer vision, 2005. iccv 2005. tenth ieee international conference on* (Vol. 2, pp. 1800–1807).

Wu, X. (2011). A linear programming approach for optimal contrast-tone mapping. *IEEE Transactions on Image Processing*, *20*(5), 1262–1272.

Wu, Z., Fuller, N., Theriault, D., & Betke, M. (2014). A thermal infrared video benchmark for visual analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)* (pp. 201–208).

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)* (pp. 818–833).

Zhang, S., Benenson, R., & Schiele, B. (2015). Filtered channel features for pedestrian detection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1751–1760).

Zhang, X., Shen, P., Luo, L., Zhang, L., & Song, J. (2012). Enhancement and noise reduction of very low light level images. In *2012 International Conference on Pattern Recognition (ICPR)* (pp. 2034–2037).

Zhao, X., He, Z., Zhang, S., & Liang, D. (2015). Robust pedestrian detection in thermal infrared imagery using a shape distribution histogram feature and modified sparse representation classification. *Pattern Recognition*, *48*(6), 1947–1960.

Zitnick, C. L., & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision (ECCV)* (pp. 391–405).

Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. In *Graphics gems iv* (pp. 474–485).