University of the Witwatersrand

School of Computer Science and Applied Mathematics
Faculty of Science

Honours Research Report

Evaluating Low-Light Image Enhancement Models Using Supplementary Virtual Image Datasets

Keaton de Jager
1636214

Supervisor(s): Dr Hairong Wang

November 2020, Johannesburg

Declaration
University of the Witwatersrand, Johannesburg
School of Computer Science and Applied Mathematics
SENATE PLAGIARISM POLICY

I, Keaton de Jager, (Student number: 1636214) am a student registered for Honours in Computer Science in the year 2020.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that ALL the work submitted for assessment for the above course is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this in not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature: 

Signed on the **30th** day of **November**, 2020 in Johannesburg.

Abstract

The realm of computer graphics research is one of the most well-funded fields of computer science. This is largely due to the ever growing popularity of video games, and their direct link to pushing the limits of computer graphics research. Now, more than ever, creating high fidelity photo-realistic images is easy. With state-of-the-art tools for ray-tracing, weather simulations, and high-end physics engines being released to the public in free-to-use applications such as *Unreal Engine 4*, it begs the question of what can we use this technology for other than video games? On the other hand, research into low-light image enhancement regularly hits a wall when it comes to collecting large numbers of real-world image pairs. The use of Computer-Aided Design software such as *Unreal Engine 4* to generate photo-realistic images under a variety of lighting conditions, may be a possible solution to this lack of data problem. This report explores how specifically designed virtual images can be used to supplement the training set of the RetinexNet low-light image enhancement model. This research is on a small scale, and results are not expected to provide a definitive answer. The results do, however, show promise. This report may serve as inspiration for further research.

Acknowledgements

I would like to acknowledge my supervisor, Dr. Hairong Wang. She had faith in me and my research idea, even though it was a risky spin on the low-light image enhancement topic. She allowed me to experiment with my research, but was always quick to help if I needed it. All while also managing the entire computer science honours degree during 2020 and the COVID-19 pandemic. I only wish I could have performed beyond her expectations.

I would also like to acknowledge my roommates and landlord for silently dealing with the sound of overworked computer fans heard throughout the night for many days in a row. Their assistance was also welcomed when I needed to get an outsiders view on the results produced, this led to some useful insights needed for the interpretation of results.

Lastly, I would like to acknowledge the the creators that built many of the assets I used from the *Unreal Engine Marketplace* to create my virtual environments. The fact that they created such useful assets that must have taken many hours and then made them freely available for use in any projects created with the engine, definitely came to my rescue when attempting to design these worlds.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Low-light image enhancement is a domain of computer vision that has been quite thoroughly explored. This is because applications of this research are far more than simply brightening an image. From object detection to visual effects and post-processing in film and TV. A robust low-light image enhancement model would prove to be a more useful tool than one might initially expect. However, there seems to be a common problem when it comes to building robust and generalizable models for low-light image enhancement. This problem is presented in many different forms but generally boils down to one common fault, a lack of realistic and diverse training data. This is especially prominent in models trained on image pairs. This report presents an experiment into a potential solution to this problem. The proposed solution takes advantage of advancements in computer graphics capabilities to generate virtual images that can be used to supplement training datasets. In this report, we explore how computer-aided design (CAD) applications can be used to produce photo-realistic images under a variety of lighting conditions, specifically designed to fill in the blanks in the training data used by prominent low-light enhancement models.

1.1 Low-Light Image Enhancement

Low-light images are surprisingly difficult to define. This is partly due to the fact that *low-light* is a very broad term. Images of many different lighting conditions are defined as low-light. Loh and Chan [2018] found this to be an important hindering aspect of most research in the field. Loh and Chan [2018] defined ten types of lighting conditions that fit into the vaguely defined *low-light* category. The ten categories are low, ambient, object, single, weak, strong, screen, window, shadow and twilight. For more information on what these categories refer to, please refer to Loh and Chan [2018]. They hope that the more specifically defined conditions would prove to be beneficial for future research into the topic as they found many of the state-of-the-art models were only focusing on one or two of these conditions. For example, a rather successful method created by Chen *et al.* [2018], focused on extreme low-light images but not any of the other nine conditions.

Loh and Chan [2018] also presented an exclusively real low-light image dataset intended to be used as a model evaluation standard. This dataset consists of over 7,000 low-light images from all ten of the lighting conditions defined in their paper. This is discussed in more detail later in this report.

1.2 Virtual Images

For the sake of removing ambiguity this report will refer to virtual images with the assumption that virtual images are images created using CAD software of some kind, also referred to as computer generated images (CGI). These images are different to synthetic images, which is a another recurring term in this report. Synthetic images refers to real world images that have been modified in some way, such as artificially darkened or corrupted by intentionally added noise and other artefacts.

A common complaint about the use of synthetic images to increase the amount of data used for training is that the images do not represent realistic lighting conditions. However, due to advancements in computer graphics research we have the ability to simulate physically accurate lighting. Another set of problems mentioned by Loh and Chan [2018] is present in papers where synthetic data was not used, but rather images were taken at different lighting conditions or with different camera settings. This is already a tedious task, but the data produced with these methods is missing some important details. These details are in the form of dynamic objects and humans. This is because the images need to be taken from the exact same location with the exact same framing and composition with no differences other than lighting. Objects in motion and people are not able to be exactly recaptured without any variation. Unless the object in motion and the people in the scene are controlled through very precise and reproducible animation engines, which is the case for generating virtual images.

1.3 Structure of this Report

This report is divided into five chapters, including this one. Chapter 2 presents background information and related work in the field of low-light image enhancement. Chapter 2 is further divided into: Section 2.1 which specifically focuses on low-light image enhancement research, Section 2.2 which focuses on virtual image generation and the research that enables these images to be a more applicable solution, and lastly Section 2.3 focuses on the use of virtual images in other fields of research such as object detection and autonomous driving.

Chapter 3 provides an overview of the experiments that were conducted over the course of this research. This chapter provides a more clearly defined research question in Section 3.1. In Section 3.2 a full break down of the low-light image enhancement model used in the experiment is provided. This section also provides more details into what considerations were made when generating the virtual images. A discussion on how results are generated and compared is also provided in this section. Chapter 4 presents the findings of the experiments conducted over the course of this research. Results are presented in two forms, quantitative quality measurements, and qualitative opinions on the visual results. Further tests were done to develop a more in depth understanding of the results and these findings as well as a discussion of the results can also be found in this section.

This report is concluded in Chapter 5, where an overall interpretation of this research is given as well as a discussion of the implications that these results may have on further research in the field.

Chapter 2

Background and Related Work

2.1 Low-Light Image Enhancement

There are two different types of low-light image enhancement models, those that use image pairs and those that do not. The former is significantly easier to build and train, however the latter has shown promising results such as the approach taken by Jiang *et al.* [2019].

When it comes to the image pair approach, significant progress has been made. There are many models that show promising results. However, as mentioned by Loh and Chan [2018], these results are not completely reliable as many of the datasets used for testing lack real world examples of low-light images. In the case of the work done by Chen *et al.* [2018], it is very difficult to compare their results to other models, as their model processes raw camera sensor information and not low-light images. To truly evaluate the performance of these models, one standardized test set of real world low-light images should be used. This is the motivation behind the creation of the ExDARK dataset [Loh and Chan 2018].

Regardless of the difficulty in conducting a fair comparative evaluation between the models, we can still discuss some of the aspects of their respective papers that may benefit from this research. The first paper that will be discussed is the model known as RetinexNet presented by Wei *et al.* [2018]. RetinexNet is based on Retinex theory, which closely models the human color perception system. It does this by breaking down an image into its reflectance and illumination. This is done by the first network that makes up RetinexNet, Decom-Net. Once the image has been decomposed into these components they are passed on to the second network that makes up RetinexNet, Relight-Net. Relight-Net serves two main purposes, improving the illumination of the illuminance map, and denoising the reflectance map. The adjusted maps are then combined using eq. (2.1), where S is the final output image (or in the decomposition process, the source image), R is the reflectance map, and I is the illuminance map [Wei *et al.* 2018].

$$S = R \circ I \tag{2.1}$$

RetinexNet was trained on a combination of two datasets. The first of which are real-world images where the low-light and normal light versions of images are captured by changing camera settings, such as ISO and exposure. The second dataset consists of synthetic image pairs. In total their training set consists of 485 real-world images and 1000 synthetic images.

The second paper that will be discussed was published around the same time as the RetinexNet paper. This model is called GLADNet and it was introduced by Wang *et al.* [2018]. GLADNet is another model trained on image pairs. In contrast to RetinexNet, GLADNet is trained on exclusively synthetic image pairs, and a much larger amount of data (roughly 5,000 image pairs). An addition to the GLADNet dataset that may be missing from the RetinexNet dataset is the inclusion of 700 grayscale images. The intention behind these images is to try to minimize any color-bias that may become an issue during training.

2.2 Virtual Images

2.2.1 Computer Graphics

In the world of computer graphics, research has sped up quite significantly in recent years. This is partly due to the increase in popularity of video games. This has motivated companies such as NVIDIA and Epic Games to invest in the development of the field. There are two main techniques in computer graphics that justify the belief that creating photo-realistic virtual images is possible. The first of which is physics based rendering (PBR), which defines how light interacts with a material. Through a combination of maps that define aspects of the material from color information to reflectivity and metalness, a designer can quite accurately simulate the realistic behaviour of light as it interacts with the material [McDermott 2018]. The second technique has been around for many years, but due to hardware limitations has not been widely implemented in many CAD applications to the same degree as in recent years. This technique is ray-tracing (or path-tracing depending on the implementation). Ray-tracing is the most physically accurate method of simulating light that is currently available. Ray-tracing is now possible, in some form or another, in real-time applications and it vastly improves the quality of shadows, reflections and overall global illumination. A perfect example of the current state of these technologies has been showcased by Engine [2020a] during the unveiling of their next generation of game engine, *Unreal Engine 5*. Unfortunately, this version of the engine will not be released until the first quarter of 2021. That being said the current version, *Unreal Engine 4*, is still incredibly powerful.

2.2.2 Unreal Engine 4

While *Unreal Engine 5* showcases the next generation of many of the techniques needed to produce some of the most photo-realistic virtual images possible, it is often thought that those features only exist in the upcoming version. This is not the case, most of these features have been rigorously tested and fine tuned in the current version of the engine. In fact, many of them have been unveiled with showcase projects made in the current version [Engine 2018a 2019a 2020b]. These showcase projects highlight the motivation behind this research. Loh and Chan [2018] state that a missing aspect of many realistic low-light image pairs is humans, and many would argue that humans are some of the most complex objects to accurately simulate. Epic Games and 3lateral [2018] showcased the *Digital Humans* project in combination with **3lateral**, which shows that creating a photo-realistic simulation of a human is now an attainable goal. A summary of this showcase is shown in Figure 2.1. Engine [2019a] was used to showcase the engines ray-tracing abilities. Both showcases are an example of the current state-of-the-art computer graphics technologies being packaged into a single

Figure 2.1: An example of the Digital Humans showcase by **Unreal Engine** and **3lateral** [Engine 2018a].

software application, that is freely available. In fact, *Unreal Engine* is not just a game engine. It is being used by a wide range of industries from video games, to architectural visualization, artificial intelligence and other fields of research, and film and TV [Engine 2018b]. An important aspect of *Unreal Engine* that should be mentioned for context as to why it was selected over all other online and offline renderers, is their partnership with **Quixel** which provides developers with unlimited, free access to the *Megascans* library. The *Megascans* library is a collection of over 10,000 photo-scanned assets and materials. This is only available for free to be used in *Unreal Engine*. In terms of researchers using the engine to create large amounts of photo-realistic images, this is a very welcome addition. A game engine is a very powerful online renderer that has more features that enable simulation of other aspects of a world. These features include powerful animation engines, sequencing tools that enable cinematics to be created with a wide range of dynamic objects, efficient physics engines [Engine 2019b] and, in some game engines, realistic weather simulations [Engine 2020b].

2.3 Other Applications of Virtual Images

Other than the different application of virtual images in the entertainment industry mentioned in the previous section, there are also a few research papers that experimented with this concept. The most notable of which was introduced by Tian *et al.* [2018]. Their paper focused on the automation of data collection for an object detection model, with specific applications in autonomous driving. They made use of virtual images to supplement existing datasets. Their motivation was due to a lack of objectively annotated data, and insufficient

variety of data in the datasets that they had available to them. They made use of another freely available game engine, *Unity*, to facilitate their experiments. They noted some of the benefits of this process included the ability to experiment with various weather and lighting conditions to generate vast amounts of data in a relatively short amount of time. In comparison to gathering this data manually in the real world, this new process was a drastic improvement. The results presented in their paper are promising and may have a significant impact on the quality of autonomous driving models as their research develops. However, their images were not rendered with visual fidelity and photo-realism as their main objective. Their focus was on generating images of objects that had been objectively annotated even in situations where annotation may not be possible without prior knowledge of the images (which is difficult if not impossible with the previous methods of manual annotation).

Chapter 3

Research Methodology

3.1 Research Question

The question that this research aims to answer is how does the performance and accuracy of a low-light image enhancement model change when specifically designed virtual images are added to the training dataset?

3.2 Experiment Design

To answer the research question presented above, three things are required. The first of which is a low-light image enhancement model that trains on image pairs. This aspect of the experiment is explored in more detail in the next section. The second thing needed is a large number of virtual image pairs. This is generally quite a laborious task as it involves very meticulously capturing thousands of images while also changing camera settings and trying not to move the camera. The third thing required is a test image dataset. Each of these aspects are needed before the experiment can be performed. The experiment itself involves training the low-light image enhancement model multiple times on various different training datasets. This requires an immense amount of time (depending on the model and the number of images). The final step in the experiment is to evaluate the images and compile these evaluations into simplified and meaningful results. Each of these aspects are explained in more detail in the sections below.

3.2.1 The Low-Light Image Enhancement Model

The initial model of choice with which these experiments would be performed was GLADNet [Wang *et al.* 2018]. However, after generating the benchmark results and beginning to train the model on the first custom dataset, this model had to be replaced by the RetinexNet model [Wei *et al.* 2018]. This last minute change was due to an oversight when picking the model. Training GLADNet on the machine available was not able to use the GPU as the GPU in the machine did not have enough memory. Training GLADNet on the CPU was projected to take 33 days per training set. Considering the number of training sets that needed to be used, this was not an acceptable training time. It is unclear what led to the training time being so poor, but the switch to RetinexNet was made and the problem fixed.

There was no significant difference between GLADNet and RetinexNet that led to the first choice being GLADNet. Both options were considered initially because both models are trained on some number of synthetic image pairs. RetinexNet was trained on two types of data, the real-world images captured by varying the camera settings to achieve low-light and normal-light versions of the same image and a synthetic dataset. The synthetic dataset was specifically designed to facilitate the model's design. This is apparent if you attempt to train the model without the synthetic images (replacing them with an equal number of image pairs from another dataset), as the results are non-existent. The model does not learn anything usable. Due to this discovery, the training sets created for this project did not alter the synthetic images set.

RetinexNet works by decomposing the input image into a reflectance map and an illuminance map. The two maps are then denoised and enhanced before being re-combined into the output image. This follows the basic idea of Retinex theory. The model is trained on RGB images of a size of 400×600 pixels and must be in *.png* image format. The network architecture for the RetinexNet model is shown in Figure 3.1. The architecture shows how the decomposition network processes both the normal-light and low-light image simultaneously. This is because the goal of the decomposition network is to produce a reflectance map from the normal-light image that is equal to the produced reflectance map from the low-light image. If these images match exactly, then it is safe to assume that the illuminance maps show only the information about the lighting in the image. The adjustment portion of the network is a made up of the Enhance-Net which is structured like a typical convolutional neural network that enhances the illuminance map, and a denoising operation that is performed on the reflectance map. The results of the adjustment layer are combined to produce the enhanced output image.

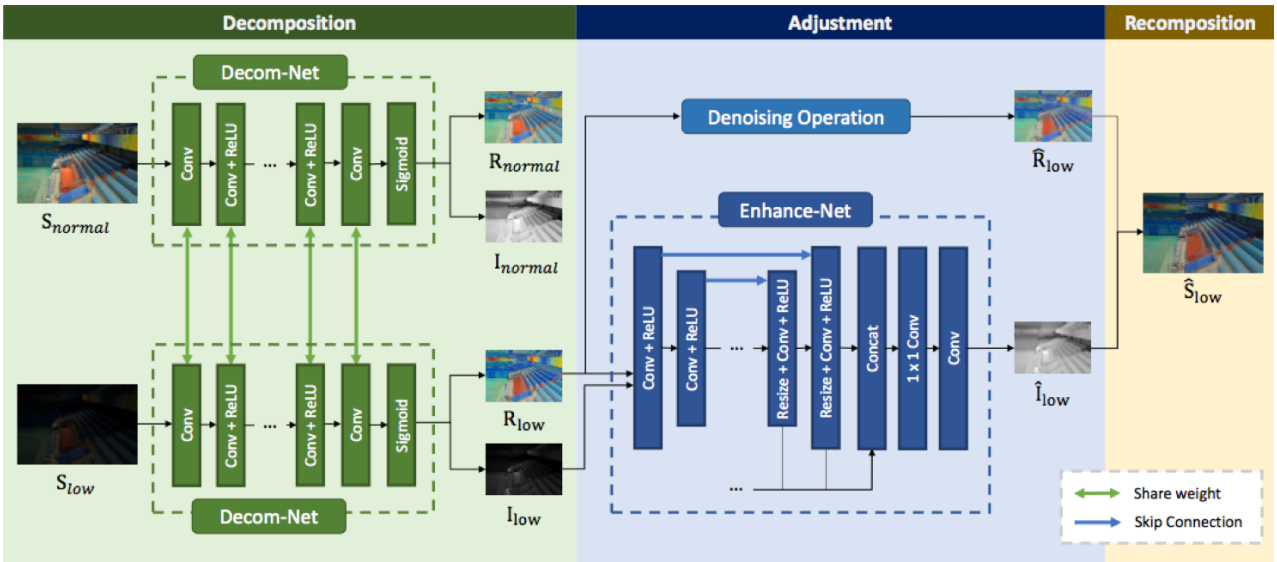

Figure 3.1: A visual representation of the RetinexNet network architecture as shown in the research paper by Wei *et al.* [2018].

3.2.2 Generating Virtual Images

The dataset consisting of entirely virtual images needed to be specifically designed to account for the shortfalls of the model, as specified in the research question. To do this the model was tested on the test dataset, and the worst performing images were analysed. However, this was done using the results produced by GLADNet, where the identified problems were as follows: (1) The enhanced version of images where strong light sources are present over exaggerated the bloom effect of the light, and (2) images that had strong contrast in the lighting conditions were particularly difficult to improve the darker spots without over exposing the brighter spots. To account for these problems image pairs were produced that consisted of the same problematic scenarios. To attempt to account for these problems, the following adjustments were made; (1) The low-light image was rendered with a reasonable amount of bloom (as a post-processing layer available in most online renderers), but the bloom level was drastically reduced in the normal-light image to reduce the over-exposed look that would become apparent, and (2) over exaggerated examples of light contrast were made in the form of scenes with very dark backgrounds and bright spot lights such as car headlights, an example of these images is shown in Figure 3.2. The switch to RetinexNet was done too late in the project to re-design the virtual images, luckily it seems the RetinexNet model has similar issues with these scenarios. It is important to note that these experiments are not taking noise into consideration. This will have an effect on the results when models are tested on real low-light images.

Figure 3.2: An example of the specifically designed high contrast image under both lighting conditions, this example also shows the adjustments made to account for bloom in the normal light image.

Additional images were made that fit in with the existing data. This data was created using free assets from the *Unreal Engine Marketplace* and *Quixel Megascans* library. Some images are rendered from showcase scenes present in either *Unreal Engine* created showcases or showcases of assets from the marketplace. A particularly appealing example of one of these scenes is shown in Figure 3.3, which made use of the architectural visualisation showcase scene. Figure 3.3 also shows the various lighting conditions that can be rendered out for the exact same image, even though the image contains dynamic objects. The animation controlling that object is perfectly reproduced at the exact same frame of the exported cinematic. The less noticeable but more important aspect of this, is that the moving object has motion blur. This is another aspect of game engines that makes them particularly useful for this application. Although this feature is not exclusive to game engines, it is likely to be enabled by default, and the default is likely to be more realistic, where other renderers might need

the designer to enable it manually, and fine-tune the effect to get to the more realistic level.

Figure 3.3: An example of the various lighting conditions at which the scene could be rendered specifically focusing on the reproducible motion of objects between each rendering. This image is of the Architectural Visualisation showcase scene [Engine 2019a].

In total 1,225 virtual images from 11 different scenes were generated. An example normal light image from each of these 11 scenes is shown in Figure 3.4. Some of the scenes were slightly more stylized than others, but their lighting was still set up the same with realism as the main focus. The aspects of the scene that were not photo-realistic exclusively consisted of certain types of scene geometry (such as buildings in a small village). Considering the lighting and material pipeline was kept consistent, it is not likely that this had an effect on the results, but it is possible that this may need further exploration.

3.2.3 Results and Evaluation Metrics

Establishing a Benchmark To establish a benchmark from which a comparison could be made, a test set needed to be created first. This was done by randomly selecting images from the ExDARK dataset. The final test set consisted of 510 low-light images from all 10 conditions as defined by Loh and Chan [2018]. The images also came from a combination of all of the object classes present in the dataset.

These images were then enhanced by the RetinexNet (and GLADNet) models using their published weights - the state of the model when it was trained on its specific dataset to produce the results presented in their respective papers. The enhanced images were then evaluated using the evaluation techniques discussed in a later section. These metrics are in the form of quality ratings, calculated using a blind image quality quantifier. These quality ratings were stored for each image (to be referenced in relation to the images later). The three best results and three worst results were identified and saved for comparison later. Finally, the average quality rating for the model was calculated, this will be needed for a simple and direct comparison between models.

Producing Results In total the model was trained five times. Each iteration of training was done with the same hyper-parameters, that match the values provided by Wei *et al.* [2018]. However, Table 3.1 shows 8 models. The first model (the benchmark) was pre-trained, and therefore was not trained again. Model 1 was trained on a slightly augmented dataset. Model 2 and Model 3 were created by combining the decomposition network and relighting network from Model 1 and the Benchmark model. These two models were evaluated in comparison to the benchmark model. This established that the slight augmentation to the dataset, reduced the performance of the decomposition network quite severely, but increased the performance of the relighting network. This was beneficial as training the of the decomposition network was far slower than the training of the relighting network. For the remainder of the models, only the relighting network was trained on the dataset, the decomposition network from the original model was used for all other experiments. This removed

Figure 3.4: An example normal light image from each of the 11 scenes used to generate the virtual data and an additional image from the architectural visualisation scene.

a variable from the experiment and allowed the results to focus on the performance changes due to one of the networks rather than both.

Model Name	No. Original Images	No. of Virtual Images	No. of Synthetic Images	Total Images
Benchmark (Decom 0, Relight 0)	485	0	1000	**1485**
Model 1 (Decom 1, Relight 1)	485	83	1000	**1568**
Model 2 (Decom 1, Relight 0)	485	83	1000	**1568**
Model 3 (Decom 0, Relight 1)	485	83	1000	**1485**
Model 4 (Decom 0, Relight 4)	0	994	1000	**1994**
Model 5 (Decom 0, Relight 5)	0	574	1000	**1547**
Model 6 (Decom 0, Relight 6)	485	391	1000	**1876**
Model 7 (Decom 0, Relight 7)	0	1225	1000	**2225**

Table 3.1: A breakdown of the composition of the training dataset used to train each model. Listed next to the name of the model in brackets is which decomposition network and which relight network was used to create the model.

Evaluation Metrics The quantitative measure used to evaluate the results from each experiment made use of a blind image quality analysis model as presented by Mittal *et al.* [2013]. The metric produced by this model is referred to as the NIQE index. The NIQE index is calculated using a complex system that is beyond the scope of this report but can be simplified down to a distance measure, where a larger NIQE index means the image is of a lower quality and a smaller NIQE index means that the image is of a higher quality. To demonstrate the type of results produced using this measure Figure 3.3 consists of the exact same image under four lighting conditions where lowest light is on the left and highest is on the right. The respective NIQE index for these images is: 9.7253, 8.7955, 7.5258 and 6.6592.

Chapter 4

Results

4.1 Quantitative Results

Once all eight of the models described in Table 3.1 had processed the 510 images in the test set, an NIQE index was calculated for each image produced by each model. Table 4.1 shows some of the more important NIQE indices in comparison to the results produced by the benchmark model. Each model has been assigned an average NIQE index. This allows for quick comparisons between models. The table specifically highlights the greatest improvement from the benchmark model to the experimental model, as well as the smallest improvement, largest decrease in quality and the average change in the ratings. A more detailed discussion of these results can be found in a later section.

Dataset	Average Rating	Largest Increase	Smallest Increase	Largest Decrease	Average Change
Benchmark	4.73611	-	-	-	-
Model 1	5.9848	8.0457	0.0064	-5.8345	-1.2488
Model 2	6.2511	7.4286	0.0080	-6.6514	-1.5150
Model 3	4.7360	0.6917	0.0001	-1.4279	+0.0001
Model 4	4.6295	3.8345	0.0030	-1.3992	+0.1066
Model 5	4.7378	3.8049	0.0017	-1.6082	-0.0016
Model 6	4.7322	2.7177	0.0001	-1.6860	+0.0039
Model 7	4.9311	4.7448	0.0010	-2.9405	-0.1950

Table 4.1: A summary of the NIQE results from each Model's test results in comparison to the benchmark model.

For the sake of context when comparing the change in performance of the models, a comparison was made between the benchmark RetinexNet model and the GLADNet model. Table 4.2 shows the improvement in the results of the GLADNet model are far greater than the decreases or improvements between the different variations of the RetinexNet model. It also shows that the GLADNet model performs better on this test set than the RetinexNet model, even though the GLADNet model was trained on purely synthetic image pairs while RetinexNet was trained on synthetic images and real-world images.

Dataset	Average Rating	Largest Increase	Smallest Change	Largest Decrease	Average Change
Benchmark	4.7361	-	-	-	-
GLADNet	3.7558	6.8784	0.0009	-1.3783	+0.9803

Table 4.2: A summary of the NIQE results from the RetinexNet benchmark model in comparison to the GLADNet model.

4.2 Qualitative Results

Figure 4.1 below shows some of the results produced by each model for the same input images. The images were selected by choosing the most recurring images from the NIQE index ratings that improved the most, performed the best or showed the greatest decrease in quality in comparison to the benchmark results. Although the best performing images from the benchmark test are also included to show how the other models performed in comparison to the best benchmarks. From initial analysis, the results show variation in color management and overall brightness, but for the most part they are quite consistent.

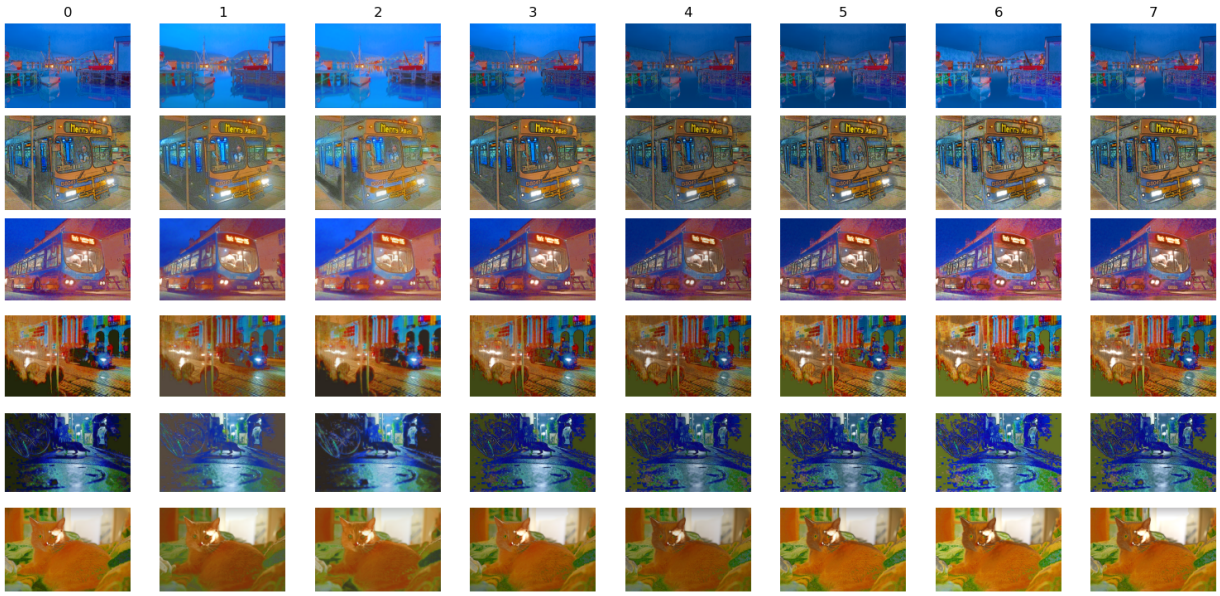

Figure 4.1: A comparison between some of the test images from each of the version of the RetinexNet model.

4.3 Insight through Biased Results

To understand the effect of training the model on exclusively virtual data, with no added artificial noise, the model had to be evaluated on low-light images that did not contain any noise. To do this, a new test set was created using the virtual images that were created for the training sets. The benchmark model and Model 7 (the model trained on the largest number of virtual images) were tested using this new test set. The new test set contained 210 virtual images of various scenes. Table 4.3 shows the summary of the comparison. The results are quite surprising, as it would make intuitive sense if Model 7 was able to out perform the benchmark model, but that is not the case. Figure 4.2 shows some handpicked results from this test in comparison to the ground truth images. It is clear that both models struggle with matching the color and saturation of the brightened image.

Dataset	Average Rating	Largest Increase	Smallest Change	Largest Decrease	Average Change
Benchmark	7.5365	-	-	-	-
Model 7	8.2163	2.1539	0.0010	2.4618	-0.6797

Table 4.3: A summary of the NIQE results from Model 7's test results on the virtual data test set in comparison to the benchmark model.

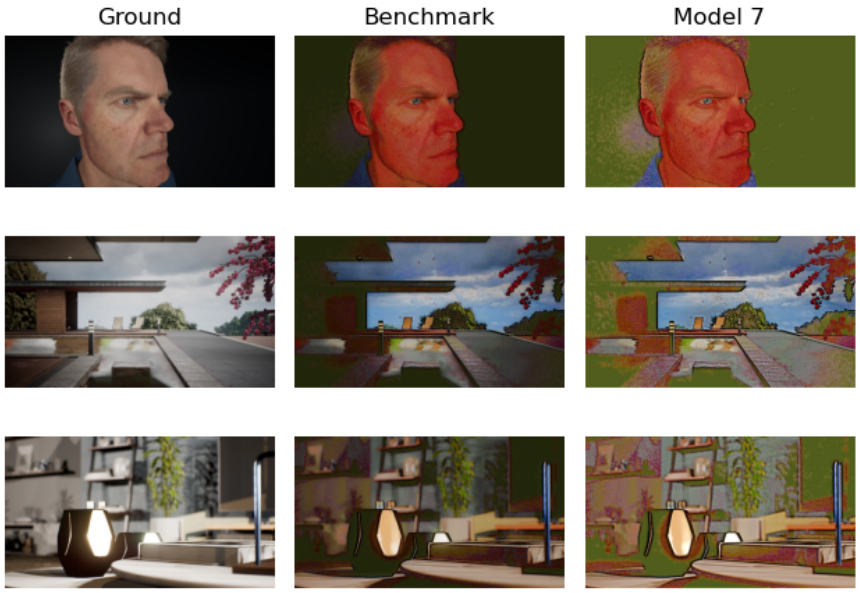

Figure 4.2: A comparison between the ground truth normal light image and the results produced by the benchmark model and Model 7.

4.4 Discussion

The quantitative results of these experiments show that there is a lot of variation in the performance of the models. A clear trend is difficult to identify. Model 4 performed the best out of all models including the benchmark, with an average NIQE index improvement of 0.1066. Model 4 was trained on 994 virtual images and no images from the original dataset other than the synthetic images. Yet both Model 7 and Model 5 which are also trained on exclusively virtual and synthetic data but in varying quantities, performed worse than the benchmark. It is clear that the decomposition network did not benefit from the virtual images as the slight addition of 83 virtual images led to a decrease in the average NIQE index of 1.5150 (which is a larger difference than the difference between the benchmark and GLADNet models).

From the quantitative results alone, it is difficult to conclude these results and declare an answer to the research question. Looking into the qualitative results, we can see that there is once again not much that stands out either for or against the virtual data. Certain combinations performed quite visibly poorly, while others are almost indistinguishable from the benchmark results. Considering these images are selected for being drastically

Figure 4.3: The input image, result from enhancement from the benchmark model, Model 4 and Model 7 for that image. Model 4 and Model 7 showed the biggest improvement from the benchmark model on this image.

different in terms of their NIQE index relative to the benchmark model, these images do not show any obvious signs of benefit or detriment from using the virtual images. It is important to note that almost all of the experiments highlighted a certain image as having the greatest improvement from the benchmark test. However, the image in question is very difficult to distinguish what caused such a drastic improvement. Figure 4.3 shows the input image, benchmark result, result from Model 4, and result from Model 7 for this commonly recurring image. It appears that the reason for this improvement in the NIQE index, is as a result of something that reduces the images visual qualitative rating. The benchmark model kept the variation of different gray values more consistently, than the other two models. This could result in the NIQE quantifier identifying the benchmark image as having a greater amount of noise than the other images. In general, the custom models seemed to handle color balance more realistically than the benchmark model. The benchmark model seemed to over-saturate the the colors picked up in the darker regions of the images, while the other models seemed to treat the colors in the entire image more equally. This led to the benchmark model producing some resulting images that were unrealistically saturated, while the other models kept a more realistically muted color. But in other cases it led to the benchmark model keeping the contrast more consistently, where the other models (such as Model 2, the first image in Figure 4.1) lose a lot of contrast and detail and appear flat.

Additionally, these experiments did not account for noise in the virtual images, or lack of noise in the images. This is in part due to the fact that both the GLADNet and RetinexNet are trained on mostly synthetic data where noise is not naturally present (although it may have been added). Regardless of whether or not the synthetic images include noise, these experiments did not remove the synthetic images from the dataset. This implies that the comparison between the models is based on the difference between real-world data and virtual data in its simplest form. If anything, this highlights a way in which the results could be improved.

Chapter 5

Conclusion

5.1 Overview and Summary

While the results presented in Chapter 4 may seem underwhelming at first, they are actually quite promising. While the model's performance fluctuated and did not consistently improve or degrade, this is a promising start to research into this topic. Consistent improvement was the goal, but considering the setbacks involved over the course of this research, the expectation was not to reach that goal. Rather, these results show that the inclusion of virtual data did not consistently reduce the performance of the model. In fact, the model's performance increased when trained on exclusively virtual and synthetic data. This may not prove that virtual images are a solution to the lack of sufficient data but it also does not prove that they are not.

This experiment was not performed under the best conditions, and the dataset created could definitely be improved. Due to a multitude of unforeseen circumstances, this experiment was not conducted at the standard that was initially expected during the proposal phase of this research. However, the project still shows promise, and the work done was still important in completing the first stage of the research endeavour. Further research is required to prove this proof of concept to be viable, but it makes intuitive sense that the idea is viable as it stands. If researchers wish to take spend more time, and money on creating a more polished virtual dataset than I was able to commit, they may see the benefits that were initially expected but not presented in this report.

There were many oversights over the course of this project that led to the underwhelming results presented in this report. Over-estimation of my ability to create these photo-realistic renderings in an engine I have never used before, was definitely a big contributing factor. My limited understanding of how to properly set up the cinematics and exports meant many hours were spent combing through the exported image sequences to remove all unusable frames. This was time spent doing something that was not accounted for in my research plan, and time that could have been spent taking more care to produce a wider variety of lighting conditions. Majority of the final image pairs are produced by changing the exposure settings on the virtual camera which matched with many of the other image pair creation techniques, but did not facilitate the exploration of adjusting the lighting itself between images. There are many different ways that the results from these experiments could be improved but there are also many ways that they could be more reliable. The main problem with their reliability, is that the virtual images selected in each version of the training set were randomly chosen from the full set. This means that some of the discrepancies in the results might be as a result of bad virtual images, as not all of them were combed out.

5.2 Implications

If this research had resulted in more promising results, the effect on research in many fields would have been substantial. Not only would it improve the efficiency of producing a robust low-light image enhancement model, but the effect would extend to any other computer vision based research. Virtual images have two major characteristics that real-world data does not have, reliable and reproducible simulations of physically accurate scenarios and the ability to re-create rare phenomenon or dangerous scenarios as accurately as possible allowing researchers to capture images from impossible perspectives. This could mean anything from simulating fog in an environment for dehazing models to train on, to simulating the aftermath of natural disaster to train robust and efficient object identification models that could be used to identify important information needed by search and rescue teams.

5.3 Conclusion

While the results of the experiments conducted in this report are not sufficient to accurately answer the research question, they also do not prohibit the further investigation of this topic. The execution of these experiments has many flaws, each of which could potentially be the cause for a lack of results. It is recommended that this idea be further researched as the implications of this concept if proven to be viable could be game changing.

Bibliography

[Chen *et al.* 2018] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2018.

[Engine 2018a] Unreal Engine. Digital humans. *Unreal Engine Documentation*, 03 2018.

[Engine 2018b] Unreal Engine. *Why Real-Time Technology is the Future of Film and Television Production*, 2018.

[Engine 2019a] Unreal Engine. Archviz interior rendering. *Unreal Engine Documentation*, Dec 2019.

[Engine 2019b] Unreal Engine. *Chaos Destruction*, Mar 2019.

[Engine 2020a] Unreal Engine. *A first look at Unreal Engine 5*, Jun 2020.

[Engine 2020b] Unreal Engine. *Sky Atmosphere*, May 2020.

[Epic Games and 3lateral 2018] Epic Games and 3lateral. *Epic Games and 3Lateral introduce digital Andy Serkis*. Epic Games, Mar 2018.

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

[Loh and Chan 2018] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Comput. Vis. Image Underst.*, 178:30–42, 2018.

[McDermott 2018] Wes McDermott. *The PBR guide:*. Allegorithmic, 2018.

[Mittal *et al.* 2013] Anish Mittal, Rajiv Soundararajan, and Alan Bovik. Making a "completely blind" image quality analyzer. *Signal Processing Letters, IEEE*, 20:209–212, 03 2013.

[Tian *et al.* 2018] Yonglin Tian, Xuan Li, and Kunfeng Wang. Training and testing object detectors with virtual images. *IEEE/CAA Journal of Automatica Sinica*, 5:539–546, 02 2018.

[Wang *et al.* 2018] Wenjing Wang, Chen Wei, Wenhan Yang, and Jiaying Liu. Gladnet: Low-light enhancement network with global awareness. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference*, pages 751–755. IEEE, 2018.

[Wei *et al.* 2018] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*. British Machine Vision Association, 2018.