Low Light Stereo Image Dataset and Enhancement

A project report

submitted by

VENKATA ANOOP SUHAS KUMAR MORISETTY

in partial fulfilment of requirements

for the award of the dual degree of

BACHELOR OF TECHNOLOGY IN

ELECTRICAL ENGINEERING

AND

MASTER OF TECHNOLOGY IN

DATA SCIENCE

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

June 2022

CERTIFICATE

This is to certify that the project titled **Low Light Stereo Image Dataset and Enhancement**, submitted by **Mr Venkata Anoop Suhas Kumar Morisetty**, to the Indian Institute of Technology Madras, for the award of the degrees of **Bachelor of Technology in Electrical Engineering** and **Master of Technology in Data science**, is a *bona fide* record of the research work done by him under my supervision. The contents of this project, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr Kaushik Mitra
Research Guide
Associate Professor
Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai 600 036

Place: Chennai

Date: June 17, 2022

ACKNOWLEDGEMENTS

I'd like to thank and express my heartfelt gratitude to my guide, Prof. Kaushik Mitra, for introducing me to the field of low-light Imaging and for his ongoing support and guidance throughout the project.

I'm especially grateful to Mohit Lamba, a PhD student in the Computational Imaging Lab, for his mentorship, guidance, and many brainstorming sessions.

I'm also grateful to Girish, a MS student in the computational imaging lab, who helped me with data collection.

I'm extremely grateful to IIT Madras for providing opportunities to explore new territories as well as an environment for personal and professional growth.

Last but not least, I'd like to thank my parents and friends for their constant support and encouragement.

ABSTRACT

KEYWORDS: Stereo, Low Light Imaging, Low light paired image dataset, Stereo low light image enhancement, Low light stereo depth estimation

Imaging in low light is challenging due to low photon count and low SNR. Short-exposure images suffer from noise and high-exposure images are frequently difficult to capture in a variety of situations. Stereo depth estimation is a critical task in many applications, including ADAS and smart phone cameras. Matching based tasks such as stereo depth estimation are highly affected by noise in low light images. As a result, having corresponding long exposure images for low exposure noisy images aids in improving depth estimation performance.With a paired stereo dataset, we can also investigate stereo low light image enhancement which subsequently improves the stereo depth estimation.

In this thesis we propose a stereo low light paired dataset consisting of 2000 short exposure stereo images, each with a corresponding long exposure stereo images. There are 200 distinct long exposure outdoor colour stereo images ,100 distinct long exposure indoor color and monochrome stereo images. Having monochrome stereo images in addition to color stereo images aids in improving the tasks like low light color monochrome image enhancement, low light color transfer, colour monochrome stereo image matching which are using simulated datasets in the literature. We also go over some of the challenges that were encountered during the dataset collection process.

Following this, We explore low-light stereo image enhancement. We propose a light-weight and fast hybrid U-net architecture guided using epipole-Aware loss that not only enhances images for visual quality but also for improving downstream depth estimation performance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Deep learning's recent success has resulted in solutions to many previously challenging problems. Computer vision, like any other field, has benefited greatly, as evidenced by many low, medium, and high-level tasks such as edge detection, segmentation, object detection, depth prediction, image enhancement, and so on. Quality of the image used as input affects the performance of all vision tasks.

Images captured in low light suffer from Poisson noise, have low grey scale, and low SNR, affecting the performance of downstream tasks and high-exposure images which are noiseless and have High SNR are frequently difficult to capture in a variety of situations. As a result, we'd like to enhance the low-light images so that they look visually appealing and have high SNR without blurring or over exposed regions.

Short Exposure (Low SNR Image) **Long Exposure (High SNR Image)**

Figure 1.1: Low Light and High light Image pair

Stereo views of scene can help in predicting the absolute depth (unlike monocular camera where depth is subjected to scale), construct 3d point clouds. Below we show a image where we superposed left and right views and we can see that nearer the objects more the disparity ,thus stereo helps in depth estimation.

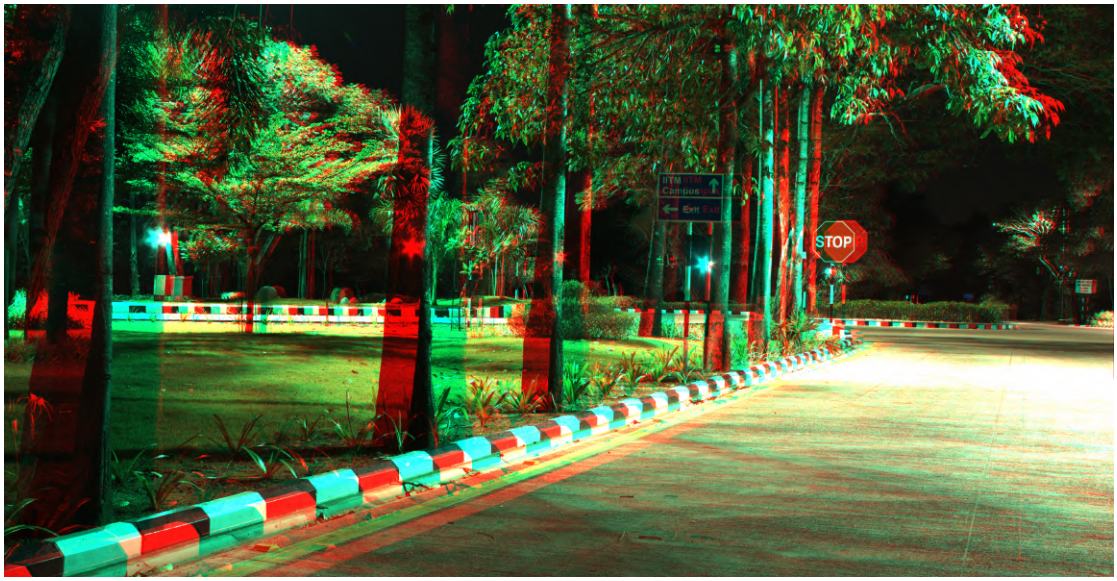


Figure 1.2: Stereo Anaglyph plot

Nighttime depth estimation is a critical task in many applications like ADAS and smart phone cameras. In low-light images, matching-based tasks such as stereo depth estimation are heavily influenced by noise. A grid of images illustrating the same is shown below.

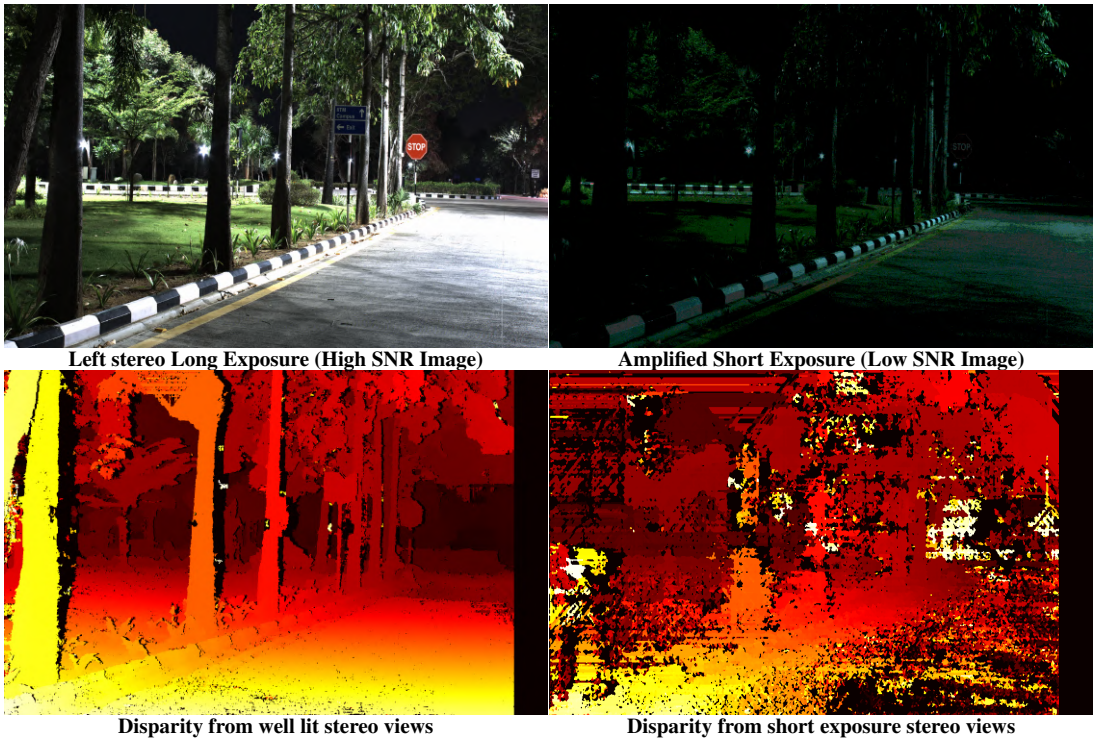


Figure 1.3: Motivation

From the above figure we see that depth from low light/exposure images are clearly sub-optimal compared to depth obtained from high exposure images. We see that there

are two possible ways to improve depth from low light images, one is to enhance the stereo pair of low light images maintaining the stereo consistency, later estimate depth from enhanced images and second is supervise the depth estimation from low light images using corresponding high light noiseless images. We see that for both possible solutions we need a paired stereo image dataset.

In this thesis ,first we propose a stereo low light paired image dataset which has 2000 low exposure exposure images, each with a corresponding high exposure images. or outdoor images, we have paired low-light and well-lit colour stereo images, and for indoor images, we additionally have paired monochrome stereo images.Having additional monochrome images along with color images, helps us to improve the tasks like low light mono color paired enhancement, low light color transfer, stereo mono color depth estimation [1] for which current methods use simulated data.

Enhancing stereo images would benefit several night-time applications that need to incorporate 3D information from the surrounding world.or example, today most self-driving cars use LiDAR to get reliable depth estimates in low-light conditions. At the same time cameras are inevitable for other ADAS-oriented tasks such as lane detection and pedestrian identification. However, if high-quality restoration can be done for low-light stereo images, the costly and bulky LiDAR may be removed for cost-efficient products. Other applications such as bokeh effect in smartphones and AR/VR headsets can similarly benefit from low-light stereo enhancement.

We propose a hybrid U-net architecture to restore dark stereo images in a way that not only benefits the visual perception of individual images but also respects the epipolar geometry for downstream applications and has low-time complexity for real-world deployment, see Fig. 5.1. To benefit visual enhancement, our hybrid architecture independently processes the left and right views in the initial scale spaces because in these scale spaces the left/right features do not align well due to large disparity. But as the scale-space becomes coarser due to repeated downsampling, the view disparity decreases, and the receptive field of the network increases significantly. So at coarser scales, we channel-wise concatenate the left and right features and let the network implicitly learn from both views. This not only favours epipole preservation but also reduces a lot of computational complexity. We avoid using 3D convolutions or attention modules to keep the network fast. Though our proposed solution is quite simple, it

is very effective and has been overlooked in the existing literature perhaps because the focus was on well-lit images.

The proposed hybrid architecture uses only 2D convolutions and so to better enforce the epipolar constraints, we train it using an Epipole-Aware differentiable loss module. The Epipole-Aware module computes the L1 loss between disparity obtained from the GT views and disparity obtained from the enhanced views. A naive approach would have used state-of-the-art depth from stereo model instead of our Epipole-Aware module to enforce the geometric constraints. But this approach has two challenges. Firstly, back-propagation through the depth estimation models is computationally very expensive requiring multiple GPUs, and thus the primary task of training the enhancement network will suffer due to memory scarcity. Secondly, almost every depth from stereo model available today has been optimized for either KITTI [2] or synthetic SceneFlow dataset [3]. Thus, they cannot be used *out-of-the-box* to train networks for images captured using any arbitrary stereo setup. Our Epipole-Aware differentiable loss module relies on classical computer vision for depth computation and has only one hyperparameter. It can thus be directly plugged in to train stereo enhancement models for any general stereo rectified setting. The Epipole-Aware loss module is not designed to replace state-of-the-art stereo depth models, which if supplied a huge amount of training data, time and memory can deliver excellent depth estimates, but to offer quick and and light-weight coarse level depth estimates sufficient for enforcing epipolar constraints during training.

1.0.1 contributions

- We propose a new novel low light paired outdoor color stereo image dataset with baseline 24cm which has 1000 short exposure stereo images with a corresponding long exposure images. It has a total 200 distinct stereo long exposure scenes.
- Aditionally,We also collected a low light paired indoor dataset of stereo color and stereo monochrome cameras seperated by 6cm which has 1000 short exposure images with corresponding long exposure image. It has 100 distinct stereo long exposure scenes.
- We explore high-speed stereo image enhancement, which although an important problem, has been largely unexplored in the existing low-light enhancement literature
- We propose an Epipole-Aware differentiable loss module which can be used *out-of-the-box* for training stereo enhancement models for any arbitrary stereo recti-

fied setting.

- Our extensive benchmarking shows that our solution is not only significantly better than most existing strategies but also offers $4-60\times$ speed-up with $15-100\times$ lower floating-point operations.

CHAPTER 2

Related works

In this section, we discuss the various datasets related to our work, which are primarily from two domains: low light image enhancement and stereo image depth estimation along with it we also mention major works in image low light enhancement and briefly discuss stereo models designed for different stereo applications.

2.1 Existing Datasets

RENOIR[4] and similar datasets were among the first to be proposed for low light image enhancement tasks. They lack a paired well-lit image that corresponds to low light for supervised training. We have the MIT-Adobe 5k dataset[5], which is similar to RENOIR[4] but is paired. The low light levels in both datasets are not extreme. Earlier several low-light image enhancement works used High Dynamic Range datasets as an alternative for paired training, such as the MEF dataset. Then [6] proposed a LOL paired dataset . Later, SID [7] proposed a dataset with paired low-light raw images and well-lit PNG images. These datasets are now used for benchmarking low-light image enhancement algorithms. There are also a few single view low light video enhancement datasets available, such as SDSD[8] and SMOID.[9]

A paired low light well lit dataset was recently proposed by MID[10]. They collected two views of the same scene, and the geometric relationship between the two views is not stereo and varies from scene to scene. This dataset cannot be used for stereo tasks because the translation varies between scenes even after stereo rectification. Aside from these, there are many unpaired low light datasets, such as Ex-dark[11], for a variety of tasks such as object detection, segmentation, and so on. However, these datasets lack a paired well-lit image.

Over time, Middlebury[12] proposed various stereo datasets with varying illumination and exposure settings. One can use low and high exposure images as paired datasets, but even low exposure images are bright and not close to low light images.

KITTI[2] and Cityscape[13] are standard stereo datasets used to testbenchmark stereo depth algorithms. However, these datasets are collected during the day. As a result, stereo depth estimation cannot be used in low light or at night. Oxford Robocar dataset[14] is a multipurpose stereo data set collected in all weather and time of day conditions to improve the adaptability of stereo depth estimation models. They also collected nighttime images, but they are noisy, blurry, and of poor quality, and they are not paired.

2.2 Low-light image enhancement

Early methods on low-light enhancement used different variations of histogram equalization to enhance the dynamic range [15, 16]. Later, people found that exploiting Retinex theory [17, 18] to decompose low-light images into illumination and reflectance components aided better enhancement [19, 20, 21]. Now-a-days, however, people use learning based networks for better low-light enhancement [22, 23, 24]. Chen *et al.* [7] proposed the famous SID dataset for extreme low-light image enhancement and the dataset has since then motivated several works on extreme low-light image enhancement [25, 26, 27, 28]. Most of these methods had a considerable computational overhead. Thus, few light-weight single image enhancement methods [29, 30] have also been proposed by slightly compromising the visual enhancement. Previous works have explored burst photography [31] for monocular image enhancement. With the success of learning based techniques for monocular image enhancement, they have been also extended for for low-light video [32, 33] and Light Field restoration [34].

2.3 Deep stereo models for different stereo applications

Stereo models have been used for wide variety of tasks such as depth estimation [35, 36, 37, 38, 39, 40], super resolution [41, 42, 43, 44, 45, 46]. Majority of depth models warp stereo features to generate a 4D cost volumes ($Height \times Width \times Disparity \times Features$) and then regress using 3D convolutions to compute disparity. Although these methods produce state-of-the-art results, using 3D convolution is computationally expensive. To alleviate this problem, recent stereo super resolution methods [41, 44]

propose relatively cheaper attention modules. While attention modules have been beneficial for well-lit images, applying them to extremely noisy low-light images may confer similar improvements. Deep stereo models have also been used for stereo deblurring [42], correcting double refraction [47], and image compression [48] but the task of light-weight enhancement of extreme low-light stereo images has been barely studied.

CHAPTER 3

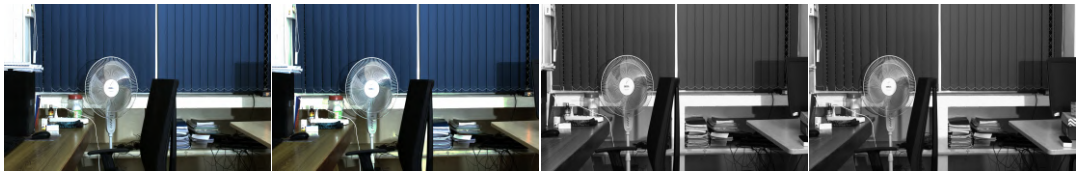
Proposed Dataset and Analysis

We gathered a novel real-world extreme low light paired stereo image dataset for training and benchmarking tasks such as low light image enhancement, low light depth prediction, and so on. This dataset is known as a **S**tereo **S**ee in the **D**ark(**SSD**). Following SID[7], we collect paired short and long exposure RAW and sRGB images for each scene.

There are 200 distinct long exposure outdoor colour stereo images captured in the beautiful IIT Madras campus and the images were captured during midnight, when illumination at the cameras is less than 10 lux for the **majority** of scenes,100 distinct long exposure indoor color and monochrome stereo images captured primarily with low artificial lighting. Long exposure stereo image pairs from the dataset are shown below.

Left View **Right View**

Figure 3.1: Sample Outdoor Image



Left colour View **Right colour View** **Left Monochrome View** **Right Monochrome View**

Figure 3.2: Sample Indoor Image

3.1 Camera Setup and Calibration

Two lightweight FLIR BFS-U3-88S6C-C machine vision cameras were used. Each sesnor produces a Bayer image with a pixel resolution of 4096x2160 and a pixel width of 3.45μm. We use a TAMRON M111FM16 lens with a focal length of 16mm, an aperture range of 1.8 to 22, and a field of view (HxV) of 42.6° x 42.6°. The cameras are controlled by the manufacturer's pySpinnaker SDK toolkit.We fixed the focus and aperture settings prior to data capture because the camera did not have an auto focus and aperture option. We settled on a focus plane of around 20m and an aperture of f/4 after extensive experiments and theoretical calculations with constraints of having Depth of field in the range of 5m-40m and an upper limit of long exposure of 12s. Below we show lens and camera

Camera **Lens** **Settings**

Figure 3.3: Camera and lens

Outdoor Setup

KITTI dataset is used for benchmarking for many outdoor stereo depth algorithms. As many stereo deep learning algorithms explicitly require disparity ranges for architecture design [35, 36, 37]. For our outdoor stereo setup, we used a 24cm baseline because it produces similar disparity/width of image ranges to KITTI [2] for a given depth range. To correct the stereo images, we need a stable and rigid stereo rig to hold the cameras and prevent any relative motion between the cameras. More on this in the 4.2 section. Our final outdoor camera setup is shown below.

Figure 3.4: Outdoor Setup

Indoor Setup

Depth range in the indoor is very less compared to outdoor. Hence we reduce the baseline 6cm. In indoor we add 2 mono chrome cameras. We have 2 color cameras seperated by 6cm in the left and 2 monochrome cameras seperated by 6cm in the right. Nearest color and monochrome cameras are seperated by 12cms. Below we show the final indoor setup.

Figure 3.5: Indoor Setup

We cannot, like any other physical system, create a perfect stereo system in which the relative pose between cameras is only translation perpendicular to the camera plane.

To correct this, we must stereo rectify the images after they have been captured. We used the built-in stereo camera calibrator app in Matlab. We capture stereo checkerboard images every day before data capture to calibrate the stereo setup's internal and extrinsic parameters. We assume that there is no relative movement between cameras within a single day of capture. The Anaglyph plot before and after rectification of images from our dataset is shown below.

Figure 3.6: Before rectification Stereo Anaglyph plot

Figure 3.7: After rectification Stereo Anaglyph plot

3.2 Protocols

To maintain consistency in the dataset, we collect data follwing specific protocols/methods as described below.Similarly to SID[7], we capture a long exposure image and a short exposure image of the scene.Long exposure times are typically between 1s and 10s.Moving objects in the scene blur long exposure images and causes misalignment of short-long exposure image pairs. As a result, we only record data from static scenes.

Our dataset's main application is image enhancement, and nighttime self-driving cars in dark or night time environments. We collect data in extremely low light conditions. Hence at each scene, we measure lux with the mobile app and primarily capture scenes with illumination at the camera is less than 10 lux.

Gamma correction is a simple post-processing technique that increases the contrast of an image.It is typically the final step in the image signal processing pipeline.As a result, we disable gamma correction during data capture and, if necessary, we can enable it during post-processing of the dataset.

To capture a low-light scene, we can use either a high exposure with a low gain/ISO or a low exposure with a high gain/ISO. A high ISO increases quantization noise and decrease the image SNR. Depth from stereo images is primarily estimated by computing matches between stereo pairs for each pixel.Noise has a negative impact on matching.This is one of the main reasons why estimating depth at night is difficult.As a result, for each long exposure image, we set the gain to zero and allow the exposure to be as long as it needs to be. Not only do we want perceptually good high exposure ground truths, but we also want the best noise performance.

Low light image enhancement using RAW images works better than sRGB images, as demonstrated by SID[7] As a result, we capture both RAW and sRGB images for each image.

For each scene, we take one long exposure GT image followed by five short exposure images. The auto exposure algorithm is used to set the long exposure, and the other five images are captured at 1/25th, 1/50th, 1/100th, 1/200th, and 1/250th fractions of the long exposure images. The camera setup is untouched between long and short exposure images. The left view images from our dataset are shown below.

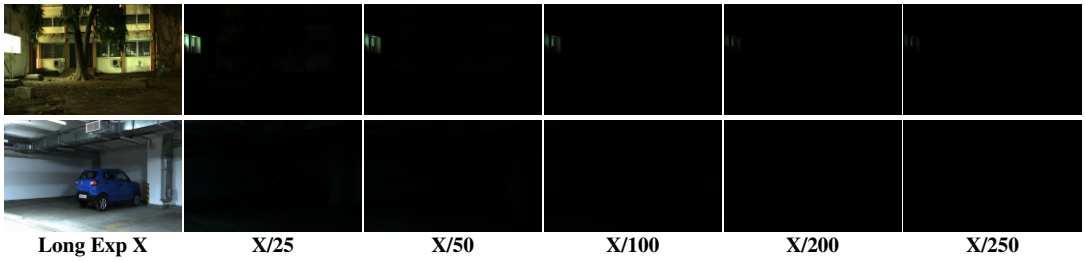


Figure 3.8: Left views

In SID[7], where outdoor scenes do not have much depth and thus are not affected by bit of wind. As the IIT Madras campus always has a light breeze passing through, there is some blur from tree leaves in few images of our dataset. Although there is a small percentage of misaligned pixels in a few images, we believe that this fraction is very small and should have no effect on the deep learning model's learning from this data.

3.3 Dataset Analysis

We collected a stereo low light paired image dataset of 200 outdoor and 100 indoor scenes.We attempted to capture scenes with a wide range of depth range, objects, lighting conditions, and exposures. In this section, we will examine the various aspects of the dataset.

3.3.1 Basic Qualitative Analysis

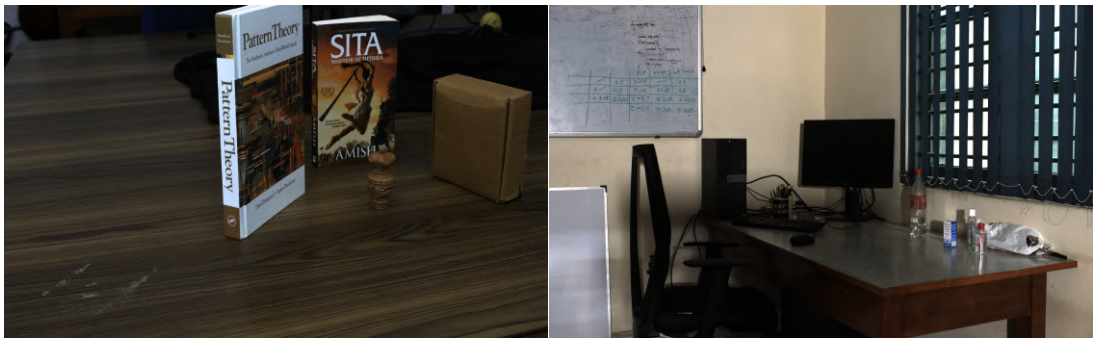
Outdoor Scenes

Our dataset was collected at night on the IIT Madras campus. The IIT Madras campus is rich in diversity, with trees, roads, buildings, vehicles, play grounds, and so on. The image grid below is filled with various scenes.

Figure 3.9: Wide spread Outdoor scenes

Indoor Scenes

In Indoor setup there are two types of scenes. One type are real scenes which are generally seen by the robot like cubicles, path, doors etc.. Other scenes are artificially created by making cameras look on the table where depth distribution will be skewed. We also show distribution of our indoor scenes in 3.11



Artificial scene **Real scene**

Figure 3.10: Indoor scenes

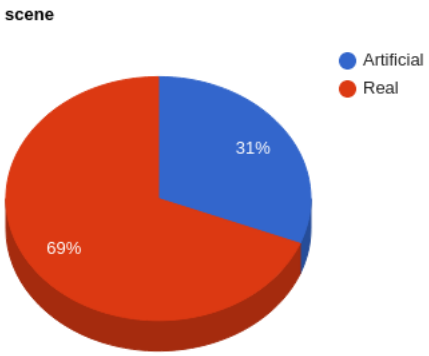


Figure 3.11: Indoor Scene Distribution

Exposure pie chart

Low light scenes have a wide range of lighting conditions, including midnight moonlight scenes, artificial street lights at various distances, and post sunset/pre sunrise scenes. The changing factor in all of the above is the amount of light that falls on the camera lens. A lux metre is typically used to measure this.During our capture, we select scenes with illumination less than 15 lux at the lens.The greater the illumination in the scene, the less exposure is required to capture the GT image.We observed that for our camera setup and settings, scenes with illumination greater than 5 lux require exposure in the range of 1 to 5s, scenes with lux less than 5 lux require exposure in the range of 5s to 10s, and scenes with lux approximately zero require exposure greater than 10s. A pie chart of exposure distribution in our dataset is shown below.

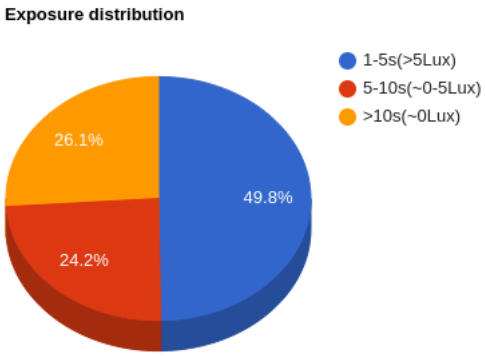


Figure 3.12: Exposure Distribution

Depth ranges with images

Current state-of-the-art deep learning algorithms for depth prediction are data-driven, and network inference results are highly dependent on training data. As a result, having scenes at different depth ranges aids the DL model's adaptability to various real-world scenarios.During our dataset collection, we attempted to collect scenes with near, medium, and far depth ranges.Images of scenes with varying depth ranges are shown below.

Near **medium** **Far**

Figure 3.13: Depth ranges

To quantify the depth distribution, we obtain disparity using [49] from well-lit ground truth images and plot a depth bar chart. We can see that the distribution is mostly concentrated in the medium and near depth ranges.We also have a sufficient number of pixels in the far depth range.

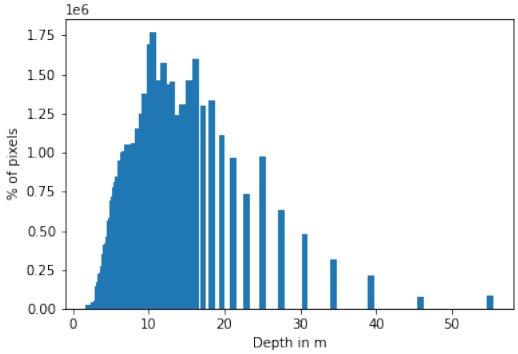


Figure 3.14: depth distribution in outdoor Images

3.3.2 Comparison with other datasets

We will compare our collected dataset to existing datasets in the low light enhancement and stereo depth research communities in this section. We compare with early low light enhancement datasets like RENOIR to stereo depth KITTI datasets. We compare on many features such as stereo, low light condition, and so on. From the below table we can clear see that only our dataset has low light stereo paired images.

Dataset	**Stereo Pair**	**Night time**	**paired**	**Outdoor**	**Indoor**	**Monochrome**
RENOIR	✗	✗	✗	✓	✗	✗
MIT-Adobe 5K	✗	✗	✓	✓	✗	✗
Middleburry	✓	✗	✓	✗	✓	✓
LOL	✗	✗	✓	✓	✓	✗
SID	✗	✓	✓	✓	✓	✗
KITTI	✓	✗	✗	✓	✗	✓
Cityscape	✓	✗	✗	✓	✗	✗
RobotCar	✓	✓	✗	✓	✗	✗
Ours	✓	✓	✓	✓	✓	✓

Table 3.1: Comparison with other datasets

3.4 Basic Experiments

In this section, we conduct basic experiments to ensure that the dataset is usable. Enhancement experiments are carried out in the thesis's following chapter.Here, we run various variations of low light stereo depth estimation experiments with the dataset.

The goal of stereo depth prediction is to determine the depth of a scene given two stereo views of the scene.In this section, we train CFNet[36], a fast and high-performing stereo depth estimation model, on the collected dataset. As shown in the following sections, We train different versions possible from our dataset as shown in following sections.

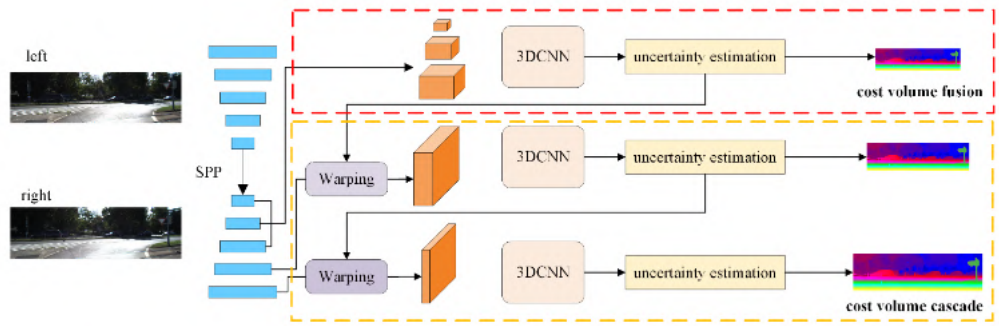


Figure 3.15: CFNet Architecture

To extract multi-scale image features from an image pair, CFnet first employs a siamese unet-like encoder-decoder architecture with skip connections. The encoder is made up of five residual blocks, followed by an SPP module to incorporate hierarchical context information more effectively.Th e multi-scale features are then divided into fused and cascade cost volume, and multiresolution disparity is predicted. Visit the CFnet paper[36] for more information on the architecture.

Loss:CFnet is trained with GT disparity supervision and employs a torch smooth L1 loss between predicted and GT disparity. However, we do not have supervised GT disparity labels in our dataset . As a result, we must train in a unsupervised manner.From [50], we use two unsupervied loss functions: photometric loss and Disparity smoothness loss.

3.4.1 Depth from GT images

This first experiment is performed to determine whether deep learning models can learn depth from our dataset's ground truth high light images. This acts as a validity check to see if the stereo constraints in the rectified images are correct and if our dataset can be used for depth problems.The CFnet's input is well-lit images, and the loss to the guide network is also from well-lit images. Because our dataset lacks depth labels, quantitative metrics are not possible.As a result, we only show qualitative results in the results section.

The results are shown when the network is trained on an RTX-3090 24GB GPU for 200 epochs from scratch, with a batch size of 4 and a learning rate of 0.0005.

3.4.2 Depth from low light images

The above experiment is repeated with the same settings except that depth is now estimated from low light images and loss is also estimated from low light images.To conduct this experiment, we used a 1/100th fraction of the GT exposure images set.

3.4.3 Depth from low light with supervision from high light images

In this experiment, we estimate depth using low-light images, but we obtain the loss to guide the network using well-lit GT images. To ensure an accurate and unbiased comparison between experiments, we use only 1/100th of the GT exposure images for training. The experimental setings are the same.

3.4.4 Results

We can see that the depth from GT is the best, which is understandable given that these are clean, noise-free images of the scenes. We can see that supervising low light images with high light images significantly improved performance when compared to supervising low light images alone. This clearly demonstrates our dataset's applicability for stereo low light depth estimation, as discussed in the introduction section.

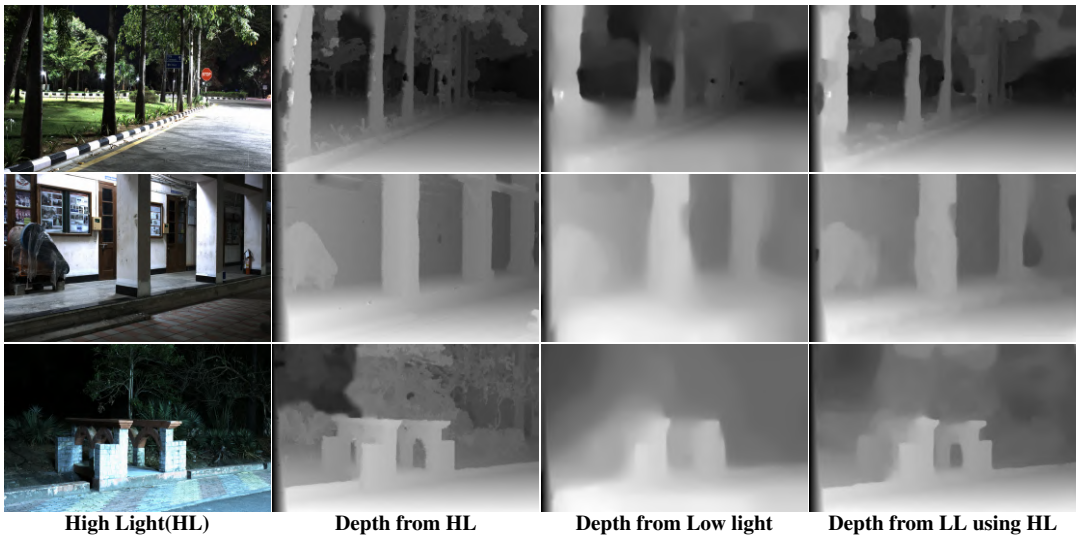


Figure 3.16: Depth Results comparison

CHAPTER 4

Challenges in Dataset collection

Collecting a real-world dataset for any task is a time-consuming and difficult process.When collecting synchronised stereo low light paired low and high exposure data, the difficulties multiply. There are numerous minor issues such as cables, scene selection, two waves of COVID, a support system to move the stereo rig, and so on, but the following are a few of the major challenges encountered during the collection process.

4.1 Static Scene

Finding a static scene with a significant depth range of 5-40m is extremely difficult due to numerous factors such as wind, moving animals obstructing the scene between capture, cars and bikes, and so on. This is a major bottleneck in our data collection process, with approximately 40 percent of scenes getting rejected or recaptured because of this reason. The figure 4.1 depict this effect.

4.2 Geometric Challenge

It is critical to have a very rigid stereo rig in order to successfully rectify images obtained from stereo cameras. We arrived at a stable rig after many iterations of improvement, and a few milestone rig designs are shown below.

4.2.1 RIG V1

For a long time, stereo images have been used for a variety of tasks. A stereo rig consists of a horizontal bar with two screws at the required distances to hold the cameras.We researched existing stereo rigs on the internet and attempted to recreate the design for our setup. Our initial setup is shown below.

Histogram equalised low light Images **Well Lit Image**

Figure 4.1: The first row shows the effect of wind causing blur, the second shows the effect of animal movement causing misalignment, and the third shows unexpected vehicle disturbance during capture.

Figure 4.2: Rig V1

On first glance, our stereo appears to be flawless, but that isn't the case. This setup is

ideal for DSLR cameras but not for our cameras. DSLR cameras have an approximate centre of mass that passes through the screw, making the entire camera very rigid and preventing any sideward motion during the capture process.Our camera lens system has a very heavy lens with the centre of mass not passing through the screw, which causes movement in both the vertical and horizontal directions of the camera, causing relative orientation R and T between cameras to change from initial calibration and scene capture. Below we show the rectified images from our setup. The rectification error is approximately in the order of 100 pixels.

Figure 4.3: Rectified Images from Rig V1

We considered several solutions to this problem.

1. Place a checkerboard in each scene and use it to estimate the R and T. R and T from a checkerboard are calculated by detecting the corners of squares. To do this correctly, the checkerboard must cover a large portion of the image. This implies that the checkerboard should be placed near the camera. Because our camera lacks autofocus, we must touch the camera at each scene, resulting in R and T changes. As a result, this solution is impractical.

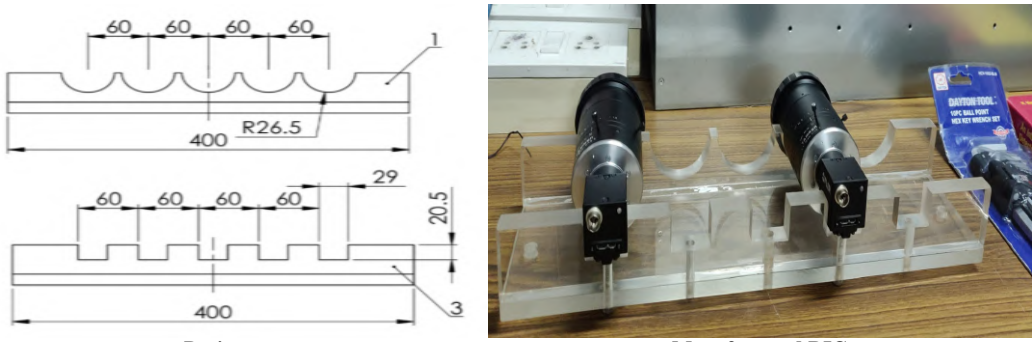
2. Compute the SIFT features and match them for each scene to estimate R and T using the 8-point algorithm. However, the SIFT features for nighttime noisy images are sparse and inaccurate to be useful for rectification.

3. The best option is to build a sturdy custom stereo rig.

4.2.2 RIG V2

To solve the problems mentioned in the preceding section, we realised that we needed to explicitly hold cameras and lenses to restrict camera movement. So we measured the dimensions of both the camera and the lens and came up with the design shown in the drawing below. We manufactured the rig with the help of the IITM workshop using the above design.

We restricted the camera's motion in the X-Y plane by surrounding it with walls, and we restricted the downward motion caused by the heavy weight lens attached to the camera. The images obtained from this rig have been rectified and are shown below. We can see that there is a significant improvement in rectification when compared to the basic mount, but the images obtained are not still perfectly stereo rectified.The rectification error is approximately 20 pixels. We are committed to improving the rig's design because we saw significant improvement in rectification error from previous design.



Design **Manufactured RIG**

Figure 4.4: RIG V2 design and product

We believe this is happening because the camera is only screwed to the rig and the lens has a cantilever mechanism. So, when the rig is mounted on a tripod and taken outside for the capture, the vibrations cause the lens to move upwards, causing a relative change in orientation between the cameras.

4.2.3 RIG V3

To address the rig's upward vertical movement in the final version. We created a counter female piece (yellow part in the figure below) to sit on top of the lens.

Aside from the major changes mentioned, as with any physical system design pro-

Figure 4.5: RIG V3

Figure 4.6: Rectified images from the final rig

cess, we have encountered numerous issues such as manufacturing defects,alignment and so on.

4.3 Temporal alignment challenge

Our dataset is stereo image dataset. Hence there should not be any temporal difference during the capture of images. There are two ways to capture the images:

1. We capture all of the left images first, followed by the right view images.Although our scenes are static, there is a slight breeze outside in some of them, causing the leaves to move around a little bit. In such cases, capturing the images sequentially may result in inconsistencies in the left and right views.For example, wind may be present during the left image capture but cease during the right image capture. As a result, we must capture both the left and right images at the same time.

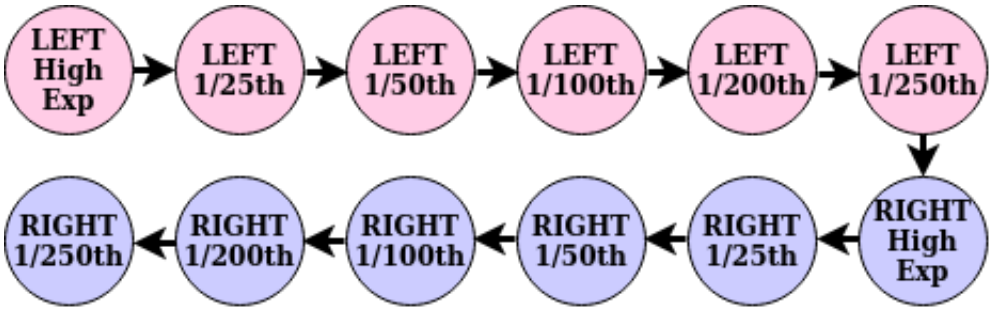


Figure 4.7: Serial mode of capture

2. As shown below, we can alternate between left and right images. However, the following process is extremely slow because we must alternate between cameras and our system must detect the camera each time.

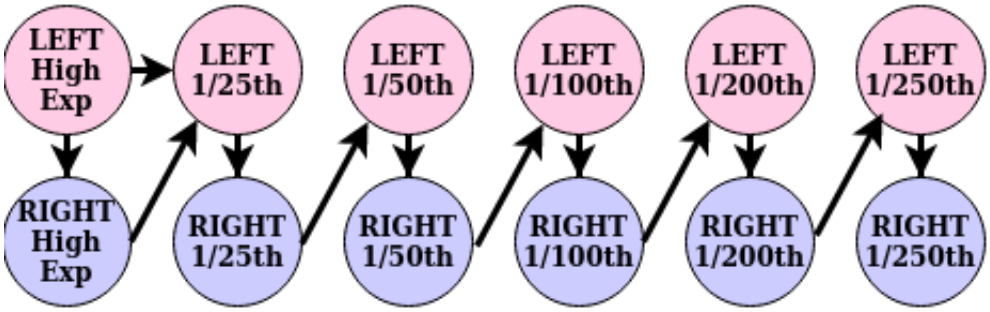


Figure 4.8: Alternating mode of capture

3.The solution is to sync the left and right capture because it does not need to alternate between cameras or go sequentially as shown below.

A GPIO port is available on our machine vision cameras. We used the triggering circuit to connect both cameras using GPIO port. When the circuit is connected, the left view becomes the primary (master) camera, and the right view becomes the secondary (slave) camera. When the Python code is executed, the primary camera sends a signal to the secondary camera, causing both cameras to begin capturing at the same time and

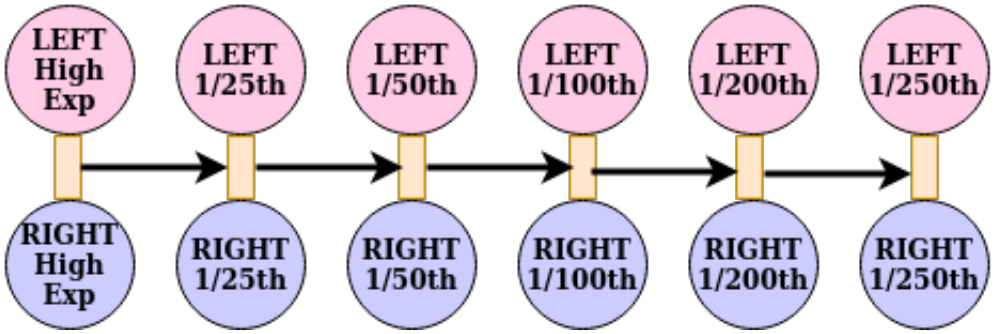


Figure 4.9: synchronous mode of capture

ensuring there is no temporal difference during capture.The circuit is shown below.

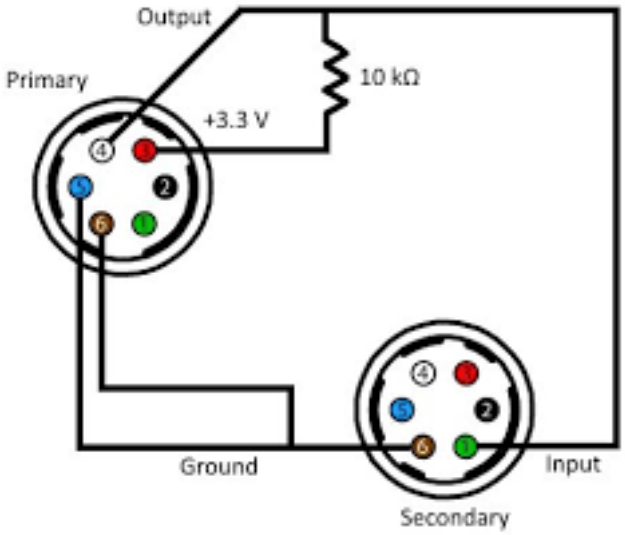


Figure 4.10: Trigger circuit

4.4 Long Exposure Dark frame Noise

We noticed a random speckle in the images we took with our cameras. This effect is very noticeable in the nighttime high exposure photos. We also noticed that as the exposure increased, so did the noise. Dark noise appears to be a common phenomenon in high exposure nighttime photography.To avoid this, we capture a dark frame immediately after the image is captured.A dark frame is an image taken with the same settings as the photo but with the cap closed.We subtract the dark frame from the image . Below we show the results on a zoomed patch from our dataset.

Note this subtraction removes the random brightening spots but creates few holes in

Figure 4.11: Dark current Noise. Zoom for better viewing experience

overexposed regions. This phenomena is more prevalent as the ISO increases.

4.5 Photometric challenge

We used 2 identical cameras for the data capture but due to differences in sensor of the cameras we are getting images of different colours. But for stereo matching we need to have images of same color.

Figure 4.12: Color Imbalance of views

1. We can solve this by including a colour calibration chart in each scene and then post-processing to obtain the appropriate colours. This works well, but it corrupts the scenes.

2. We discovered that the Y channel of the captured images matches, indicating that the problem is limited to colour matching.As a result, we considered computing a colour correction matrix.Through SIFT matches, we can learn the colour correction matrix.As previously stated, SIFT matching is ineffective at night.This process also results in a global transformation, which incorrectly corrects dark and overexposed regions.

3. We used the photoshop matching white balance option to colour correct the image after realising we needed to the local transformations. We see that the results of

this method are positive. but we don't know algorithm it uses.

4.Finally we got our best resuls using histogram matching because it uses cumulative distribution to find a local transformation function for different intensity ranges and matches distribution between left and right images.The results of histogram matching are shown below. Scikit-image histogram matching was used. We see that still there is scope of improvement in color matching which can be explored later.

Left View **color corrected right View**

Figure 4.13: Histogram matched views

CHAPTER 5

Proposed Stereo Enhancement algorithm

5.1 Dataset simulation

There is no publicly available dataset on extreme low-light stereo enhance- ment for quantitative benchmarking. Thus following previous works [51, 52, 53], which faceda similar challenge , we transformed KITTI2015 and cityscapes well-lit stereo images into low-light images. But instead of naively adding Gaussian noise and darkening the image using gamma functions we follow a more principled approach for a realistic modelling [54] as described follows.

In real low-light conditions, the noise affects the image quality right from the point of image acquisition by the sensor. After the sensor acquisition, the image undergoes several transformations such as tone mapping, white balancing and gamma decompression to produce the final RGB image or more specifically the sRGB image. Thus, to transform well-lit camera images into low-light images we do not follow the crude approximation of simply darkening the sRGB image and adding White Gaussian noise. We rather follow the pipeline proposed in [54] to first transform the well-lit sRGB images into demosaiced Linear RGB images, then add the heteroscedastic noise and finally convert them back to sRGB images.

5.2 Network Architecture

We design a hybrid U-net architecture, that accepts a pair of low-light stereo images and outputs the restored stereo views. It is designed to enable each view to harness the information present in the corresponding stereo view with- out using any computationally expensive 3D convolutions or attention mech- anisms. Fig. 1 shows our proposed network, which operates at 6 scale spaces: 1, 1/2, 1/4, 1/8, 1/16 and 1/32th resolution of the input image. In the initial few scale spaces the stereo features do not align due

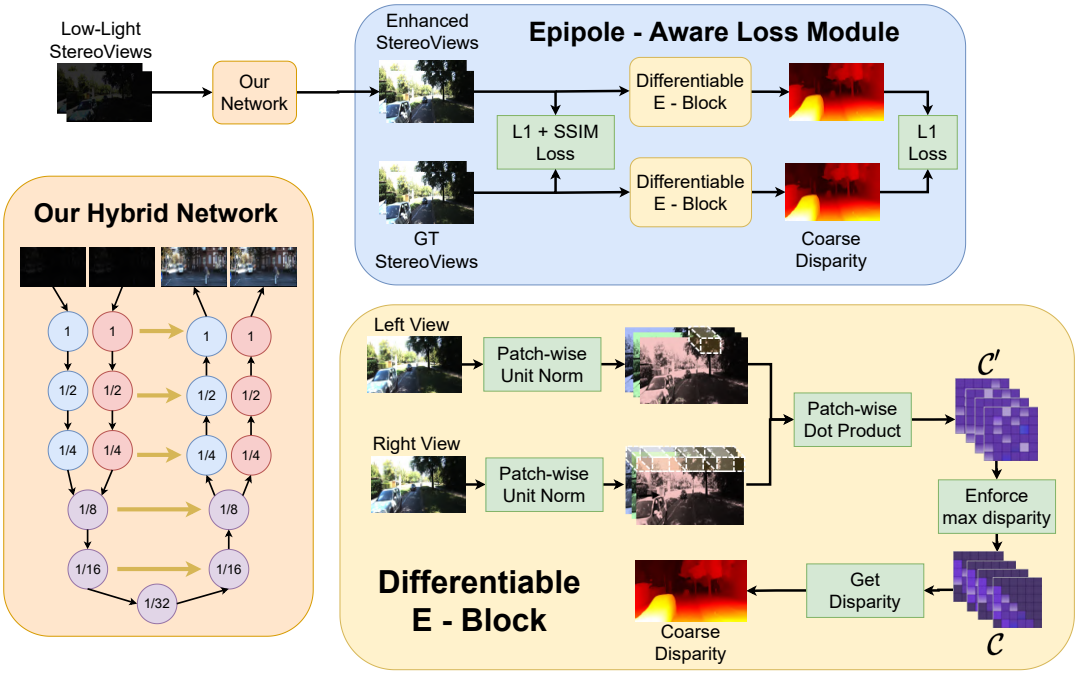


Figure 5.1: Our hybrid U-net architecture restores extreme low-light stereo images while preserving the epipolar geometry. It delivers high inference speed because we only use 2D convolutions and enforce the epipolar constraints by training it using the Epipole-Aware loss module. Our hybrid architecture further reduces the computational overhead by jointly processing the stereo features at lower resolutions which is possible due to the network's large receptive field and the reduced disparity between stereo features.

to large disparity. We thus process them independently by running the convolutional kernels twice, once for each stereo feature. But as the feature dimensions reduces, due to repeated down- sampling operations, the misalignment between the stereo features also reduces and the network's receptive field increases. For example, the maximum pixel dis- parity in the KITTI [2] dataset is between 200 to 256 and for CityScape [13] it is even lower. So at the 1/8th resolution scale space, the maximum pixel disparity will be 25 to 32. But our network's receptive field just after the first convolu- tional layer in this scale space is 36×36. Thus, at 1/8th resolution scale space we channel-wise concatenate the stereo features and process them jointly for the remaining scale spaces. This not only facilitates the exchange of information between the stereo features but also avoids doing repeated convolutions. To save computations, we allot more convolutional kernels to later scale-spaces and do not use too many convolutional kernels in the initial scale spaces.

5.3 Depth loss

Our hybrid U-net uses only 2D convolutions to achieve high-speed inference. We thus train it using the Epipole-Aware Loss Module to better enforce the epipolar constraints. The module has two components, namely the photometric loss denoted by $\mathcal{L}_{ph}$ and the disparity consistency loss denoted by $\mathcal{L}_{disp}$. $\mathcal{L}_{ph}$ computes the L1+SSIM loss between enhanced stereo views and the ground-truth (GT) stereo views. $\mathcal{L}_{disp}$ on the other hand computes the L1 loss between disparity obtained from the enhanced views and disparity obtained from the GT stereo views. The overall loss function $\mathcal{L}$ can be thus summarised as:

$$\mathcal{L} \quad = \quad \mathcal{L}_{ph} + \lambda \cdot \mathcal{L}_{disp} \tag{5.1}$$

Photometric loss: Let $\mathbf{L}_{en}$ and $\mathbf{R}_{en}$ denote the enhanced left and right stereo views, and $\mathbf{L}_{GT}$ and $\mathbf{R}_{GT}$ denote the GT stereo views. Further, let $L_1(\cdot,\cdot)$ compute the difference between l_1 norm of the input tensors and $d_{ssim}(\cdot,\cdot) = 1 - SSIM(\cdot,\cdot)$. The photometric loss is thus computed as,

$$\begin{aligned} \mathcal{L}_{ph} \quad &= \quad 0.5 \cdot [L_1(\mathbf{L}_{GT}, \mathbf{L}_{en}) + L_1(\mathbf{R}_{GT}, \mathbf{R}_{en})] \\ &\quad + 0.5 \cdot [d_{ssim}(\mathbf{L}_{GT}, \mathbf{L}_{en}) + d_{ssim}(\mathbf{R}_{GT}, \mathbf{R}_{en})] \end{aligned} \tag{5.2}$$

Differentiable E - Block: Given a pair of stereo rectified views $\mathbf{L} \in \mathbb{R}^{H \times W \times 3}$ and $\mathbf{R} \in \mathbb{R}^{H \times W \times 3}$, the E - Block computes the disparity between the two views in a way that allows back-propagation through it.

Given any pixel in the left view $\mathbf{L}$ we construct a $M \times M$ (in this work $M = 31$) patch around it and compute a unit normalised dot product with every other patch of same dimension in the right stereo view along the horizontal epipolar line. In this way we construct $\mathcal{C}' \in \mathbb{R}^{H \times W \times W}$, such that each entry $\mathcal{C}'_{i,j,k}$ in $\mathcal{C}'$ is computed as:

$$\mathcal{C}'_{i,j,k} \quad = \quad \frac{\mathbf{LP}}{||\mathbf{LP}||_1} \bullet \frac{\mathbf{RP}}{||\mathbf{RP}||_1} \; \forall \; i \in [1, H] \text{ and } \; \forall \; j, k \in [1, W] \tag{5.3}$$

where, $\mathbf{LP} \in \mathbb{R}^{M \times M \times 3}$ is a patch around the pixel (i, j) in $\mathbf{L}$ and $\mathbf{RP} \in \mathbb{R}^{M \times M \times 3}$ is

a patch around the pixel (i, k) in $\mathbf{R}$. However, computing $\mathcal{C}'_{i,j,k}$ can be computationally expensive and we prefer a lighter operation to not compromise the training of our hybrid U-net architecture. For example, if we ignore the unit normalisation step for simplicity, each $\mathcal{C}'_{i,j,k}$ requires at least $3M^2$ multiplications. We however found that more than the size of the chosen path, it is the context which matters more. We thus introduce a dilation term d, wherein every d^{th} rows and columns of the patch are considered to compute $\mathcal{C}'_{i,j,k}$. This way only $3\left(\frac{M}{d}\right)^2$ multiplications are required. In this work we set $d = 3$ and experiment with this idea during the ablation studies.

$\mathcal{C}'$ now has all the information required for computing a quick and coarse level disparity. The disparity for any pixel (i, j) in $\mathbf{L}$ is computed as,

$$j - k', \text{where } k' = \underset{k}{argmax}(\mathcal{C}'_{i,j,k}) \tag{5.4}$$

Though $argmax$ conventionally does not allow back-propagation, we make it differentiable by forcing the gradients through (i, j, k') in $\mathcal{C}'$ as 1 and all other gradient to 0. This simple workaround is also sometimes used to make the *maxpooling* layer differentiable. More sophisticated methods like SGM [49] additionally enforce smoothness constraints which definitely help in obtaining finer disparities. But for the extreme low-light enhanced views, very fine textures are hard to recover and so we found coarse level disparity good enough to train our network. This not only avoids the challenges involved in making additional constraints differentiable but also keeps the operation computationally light.

Once we have computed the disparity, we also compute a confidence map $\mathbf{C} \in \mathbb{R}^{H \times W}$ as follows,

$$\mathbf{C}_{i,j} = \begin{cases} 1 & \text{if } \mathcal{C}'_{i,j,k'} \geq \Psi \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

where Ψ is the mean of all the entries in $\mathcal{C}'$ where *argmax* in Eq. 5.4 was obtained.

For most stereo setup we have a fair amount of idea on the maximum disparity, $disp_{max}$, required because information such as baseline and camera focal length is gen-

erally known. We incorporate this information to regularise $\mathcal{C}'$. Specifically, we define a new tensor $\mathcal{C} \in \mathbb{R}^{H \times W \times W}$,

$$\mathcal{C}_{i,j,k} = \begin{cases} \mathcal{C}'_{i,j,k} & \text{if } 0 \leq j-k \leq disp_{max} \\ \text{invalid} & \text{otherwise} \end{cases} \quad \forall i \in [1,H] \text{ and } j,k \in [1,W] \tag{5.6}$$

and then use only the valid entries in $\mathcal{C}$ to compute the disparity and the confidence. For all our experiments we only vary this single hyperparameter $disp_{max}$. Further, in the ablation studies we show that in rare cases when $disp_{max}$ is not known, it can be set to $disp_{max} = \infty$ and the predicted disparity is still quite good.

Disparity Consistency Loss: Let, $\mathbf{D}_{en}$ and $\mathbf{C}_{en}$ be the disparity and confidence map produced by the E - Block for the enhanced stereo views. Likewise, let $\mathbf{D}_{GT}$ and $\mathbf{C}_{GT}$ be the disparity and confidence map for the GT stereo views. The disparity consistency loss, $\mathcal{L}_{disp}$ is thus computed as,

$$\mathcal{L}_{disp} = L_1(\mathbf{D}_{GT} \cdot \mathbf{C}_{en} \cdot \mathbf{C}_{GT}, \mathbf{D}_{en} \cdot \mathbf{C}_{en} \cdot \mathbf{C}_{GT}) \tag{5.7}$$

CHAPTER 6

Experimental results

6.1 Simulated Data results

Method	Perceptual		Depth		Inference Speed		
	PSNR↑	SSIM↑	RMSE↓	D1%↓	CPU(s)↓	GPU(ms)↓	GFLOPs↓
SID [7]	16.32	0.696	7.78	22.2	2.23	<u>30.49</u>	<u>200.50</u>
SGN [26]	16.30	0.692	7.93	22.2	<u>2.10</u>	30.86	203.92
StereoSR* [46]	20.97	0.664	7.91	20.2	12.17	147.50	1101.86
PASSR* [41]	14.86	0.699	6.69	<u>18.9</u>	38.31	473.82	2301.25
DASSR* [42]	20.09	0.673	8.12	23.9	6.40	132.27	407.62
CFNet* [36]	<u>24.56</u>	<u>0.718</u>	<u>7.37</u>	21.9	15.75	312.33	1278.44
Ours	**25.16**	**0.726**	**5.70**	**17.7**	**0.35**	**7.47**	**13.12**

Table 6.1: Quantitative comparisons on the KITTI dataset. The best scores are in **bold** and second best <u>underlined</u>. Our method achieves the best performance on all metrics while delivering real-time inference speed.

In 6.1 we benchmark our method on 7 metrics: PSNR and SSIM for compar- ing visual enhancement; RMSE and D1 bad pixel percentage [2] for comparing depth computed from enhanced views; and CPU time, GPU time and Floating- Point Operations (FLOPS) for measuring inference computational complexity. For measuring computational overhead we considered the time/operations re- quired to enhance both left and right views with full spatial resolution. We find that our method performs significantly better than most methods while exhibiting real-time inference speed, necessary for real-world applications. These quantitative results are also supported by the qualitative results shown in Fig. 3. For computing RMSE and D1 metric we have used the

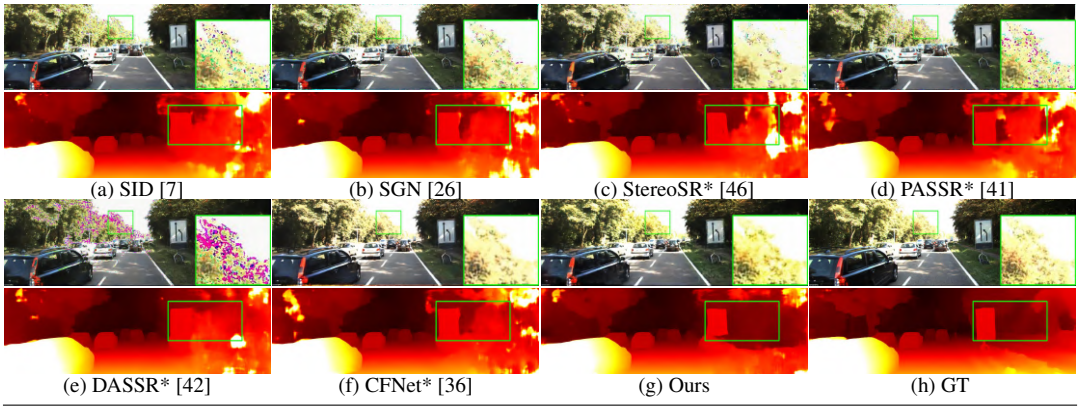


Figure 6.1: The figure shows the left view enhanced by different methods, and the depth computed using the low-light enhanced stereo views. Our method performs significantly better than most methods. With respect to CFNet*, our visual results are comparable but with $40\times$ higher inference speed.

LiDAR GT available in the KITTI dataset. But as LiDAR outputs semi-dense depth, for visual comparison in Fig. 3 we have shown the dense depth map obtained from GT stereo views. In general, we observe that stereo methods perform better than SID and SGN monocular methods. This is expected as monocular methods do not benefit from the corresponding views. Further, SID uses max-pooling for downsampling which suffers from gradient sparsity and transposed convolution for upsampling, which has been reported to lower the performance [55]. Stereo models like PASSR* and StereoSR* do feature matching for final restoration. Contrary to their approach CFNet* uses 3D convolutions for enhancement and thus achieves the best results compared to existing stereo models. This is because, using attention modules or

feature correlation is beneficial for well-lit images but not for extremely low- light images having poor contrast and large amount of noise. This fact is also evident from Fig. 3, where the enhancement done by all previous methods except for CFNet* and ours, suffers from the 'Halo Artifacts' resulting from incorrect color restoration in the small vicinity of saturated pixels [56, 19, 57]. This is the main reason for superior PSNR of CFNet* and our method. Finally, CFNet* and DASSR* estimate intermediate disparity to warp the views. We, however, do not leverage such ideas in our model because view warping using disparity computed from intermediate low-light features is prone to errors. We rather simply channel-wise concatenate the features and let the network implicitly learn using 2D convolutions by enforcing epipolar constraints using the Epipole-Aware loss module over enhanced views. Doing so not only helps our method achieve better visual enhancement and depth estimates than all previous methods in

both Tab. 1 and Fig. 3 but also keeps the memory footprint low.

6.2 Real Data Results

As mentioned earlier in the thesis, we collected a real light stereo paired dataset. As we only showed results on simulated data, here to further check the applicability on real data we show the results and both enhancement and subsequent disparity are fairly good. We are thus quite confident that our n/w generalises to real data as well.

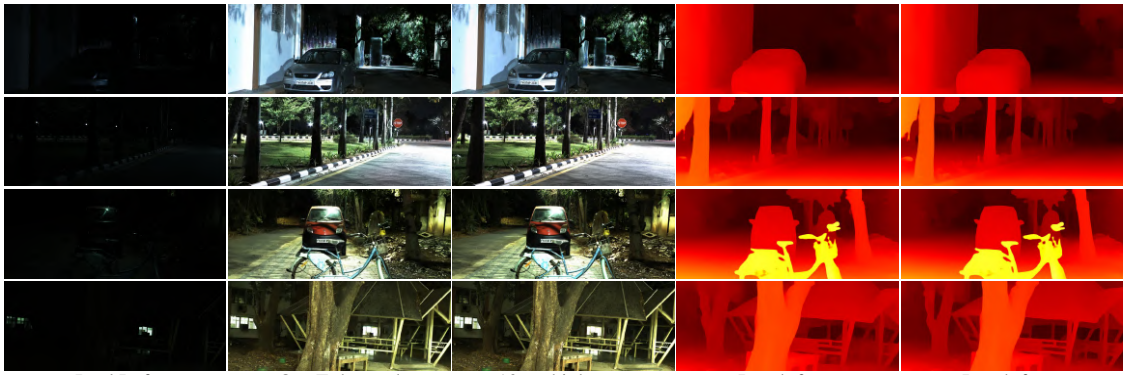


Figure 6.2: Real low light data results

6.3 Abalation studies

	Independent Feature Extraction	Our Feature Extraction	λ $\mathcal{L}_{disp}$ weight	Perceptual (PSNR)	Depth (RMSE)
Net-I	✓	✗	0	**25.26**	7.82
Net-II	✗	✓	0	25.17	6.00
Net-III	✗	✓	1	24.90	**5.38**
Proposed	✗	✓	0.1	25.16	5.70

Table 6.2: Ablation Study on the proposed method using the KITTI dataset. Our style of feature extraction benefits epipolar constraints while only slightly lowering the PSNR for visual enhancement. The table also shows the trade-off between perceptual enhancement and depth estimation.

Tab. 6.2 reports the quantitative comparison for stereo enhancement by re-training different versions of our method on the KITTI dataset. Results on Net-I and Net-II demonstrates the benefit of our hybrid architecture. For Net-I throughout our U-net the features for left and right views were processed independently. For Net-II our hybrid

style of feature extraction, as shown in Fig. 5.1, was used. Compared to Net-1, the depth prediction metric is much better by 1.82 units while experiencing only a tiny drop of 0.09 dB PSNR. Moreover, Net-II also has computational advantages. For example, for a 2MP image Net-I requires 77.08 GFLOPS while Net-II only requires 57.42 GFLOPS. We next train Net-II by including our disparity consistency loss, $\mathcal{L}_{disp}$ with weightage of $\lambda = 1$. This improves the depth metric RMSE but lowers the PSNR. This perception-depth trade-off was also noticed in [42]. Since we wish to have both good enhancement as well as good depth, we choose $\lambda = 0.1$.

CHAPTER 7

Conclusion

In this thesis, we propose a stereo low light paired dataset consisting of 2000 short exposure stereo images, each with a corresponding long exposure stereo images. There are 200 distinct long exposure outdoor colour stereo images ,100 distinct long exposure indoor color and monochrome stereo images. We faced numerous challenges while collecting the dataset, which necessitated solutions from both software and hardware domains. With basic experiments, we also demonstrated qualitatively that our hypothesis of improving night time depth estimation can be done with our dataset.

Following this, we address extreme low-light stereo enhancement which has been almost unexplored in low light image enhancement literature. We proposed a hybrid U- net architecture which faithfully restores the stereo images belonging to various datasets, while preserving the epipolar geometry. The inference speed of our network is much better than existing stereo methods because we use only 2D convolutions and enforce the epipolar constraints during training by using the epipole-aware loss module. We showed that this module can be used out-of- the-box for training on different types of datasets such as KITTI and our proposed dataset.Our network is close to real-time and offers 4 to 60× speedup with 15 to 100× lower floating-point operations compared to existing strategies.

REFERENCES

[1] H.-G. Jeon, J.-Y. Lee, S. Im, H. Ha, and I. S. Kweon, "Stereo matching with color and monochrome cameras in low-light conditions," in *CVPR*, 2016, pp. 4086–4094.

[2] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *CVPR*, 2016, arXiv:1512.02134. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2016/MIFDB16

[4] J. Anaya and A. Barbu, "Renoir - a dataset for real low-light noise image reduction," *arXiv preprint arXiv:1409.8230*, 2014.

[5] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input / output image pairs," in *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[6] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *BMVC*, 2018.

[7] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *CVPR*, 2018.

[8] R. Wang, X. Xu, C.-W. Fu, J. Lu, B. Yu, and J. Jia, "Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9680–9689.

[9] H. Jiang and Y. Zheng, "Learning to see moving objects in the dark," 10 2019, pp. 7323–7332.

[10] W. Song, M. Suganuma, X. Liu, N. Shimobayashi, D. Maruta, and T. Okatani, "Matching in the dark: a dataset for matching image pairs of low-light scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6029–6038.

[11] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Computer Vision and Image Understanding*, vol. 178, pp. 30–42, 2019.

[12] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1, 2003, pp. I–I.

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016, pp. 3213–3223.

[14] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017. [Online]. Available: http://dx.doi.org/10.1177/0278364916679498

[15] Y.-T. Kim, "Contrast enhancement using brightness preserving bi-histogram equalization," *IEEE transactions on Consumer Electronics*, vol. 43, no. 1, pp. 1–8, 1997.

[16] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.

[17] E. H. Land and J. J. McCann, "Lightness and retinex theory," *Josa*, vol. 61, no. 1, pp. 1–11, 1971.

[18] E. H. Land, "The retinex theory of color vision," *Scientific american*, vol. 237, no. 6, pp. 108–129, 1977.

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

[20] X. Fu, D. Zeng, Y. Huang, Y. Liao, X. Ding, and J. Paisley, "A fusion-based enhancing method for weakly illuminated images," *Signal Processing*, vol. 129, pp. 82–96, 2016.

[21] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *CVPR*, 2016.

[22] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *CVPR*, 2020.

[23] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *CVPR*, 2020.

[24] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *CVPR*, 2019.

[25] P. Maharjan, L. Li, Z. Li, N. Xu, C. Ma, and Y. Li, "Improving extreme low-light image denoising via residual learning," in *Int. Conf. Multimedia and Expo (ICME)*, 2019.

[26] "S author=Gu, Shuelf-guided network for fast image denoising,hang and li, yawei and gool, luc van and timofte, radu," in *ICCV*, 2019.

[27] K. Xu, X. Yang, B. Yin, and R. W. Lau, "Learning to restore low-light images via decomposition-and-enhancement," in *CVPR*, 2020.

[28] Y. Atoum, M. Ye, L. Ren, Y. Tai, and X. Liu, "Color-wise attention network for low-light image enhancement," in *CVPR Workshop*, 2020.

[29] M. Lamba, A. Balaji, and K. Mitra, "Towards fast and light-weight restoration of dark images," in *BMVC*, 2020.

[30] L. Mohit and M. Kaushik, "Restoring extremely dark images in real time," in *CVPR*, 2021.

[31] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, "Burst denoising with kernel prediction networks," in *CVPR*, 2018.

[32] C. Chen, Q. Chen, M. N. Do, and V. Koltun, "Seeing motion in the dark," in *ICCV*, 2019.

[33] H. Yue, C. Cao, L. Liao, R. Chu, and J. Yang, "Supervised raw video denoising with a benchmark dataset on dynamic scenes," in *CVPR*, 2020, pp. 2301–2310.

[34] M. Lamba, K. K. Rachavarapu, and K. Mitra, "Harnessing multi-view perspective of light fields for low-light imaging," *IEEE Transactions on Image Processing*, vol. 30, pp. 1501–1513, 2021.

[35] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *CVPR*, 2018.

[36] Z. Shen, Y. Dai, and Z. Rao, "Cfnet: Cascade and fused cost volume for robust stereo matching," in *CVPR*, 2021.

[37] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *ICCV*, 2017.

[38] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," *NIPS*, vol. 33, pp. 22 158–22 169, 2020.

[39] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *CVPR*, 2019, pp. 3273–3282.

[40] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," *arXiv preprint arXiv:1709.00930*, 2017.

[41] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo, "Learning parallax attention for stereo image super-resolution," in *CVPR*, 2019, pp. 12 250–12 259.

[42] B. Yan, C. Ma, B. Bare, W. Tan, and S. Hoi, "Disparity-aware domain adaptation in stereo image restoration," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 176–13 184.

[43] X. Ying, Y. Wang, L. Wang, W. Sheng, W. An, and Y. Guo, "A stereo attention module for stereo image super-resolution," *IEEE Signal Processing Letters*, vol. 27, pp. 496–500, 2020.

[44] Y. Wang, X. Ying, L. Wang, J. Yang, W. An, and Y. Guo, "Symmetric parallax attention for stereo image super-resolution," in *CVPR workshop*, 2021, pp. 766–775.

[45] Q. Xu, L. Wang, Y. Wang, W. Sheng, and X. Deng, "Deep bilateral learning for stereo image super-resolution," *IEEE Signal Processing Letters*, vol. 28, pp. 613–617, 2021.

[46] D. S. Jeon, S.-H. Baek, I. Choi, and M. H. Kim, "Enhancing the spatial resolution of stereo images using a parallax prior," in *CVPR*, 2018, pp. 1721–1730.

[47] H. Kim, A. Meuleman, D. S. Jeon, and M. H. Kim, "High-quality stereo image restoration from double refraction," in *CVPR*, 2021.

[48] X. Deng, W. Yang, R. Yang, M. Xu, E. Liu, Q. Feng, and R. Timofte, "Deep homography for efficient stereo image compression," in *CVPR*, 2021.

[49] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE TPAMI*, vol. 30, no. 2, pp. 328–341, 2008.

[50] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6602–6611.

[51] T. Remez, O. Litany, R. Giryes, and A. M. Bronstein, "Deep Convolutional Denoising of Low-Light Images," *arXiv:1701.01687*, 2017.

[52] S. Park, S. Yu, B. Moon, S. Ko, and J. Paik, "Low-light image enhancement using variational optimization-based retinex model," *IEEE Trans. Consumer Electronics*, vol. 63, no. 2, pp. 178–184, 2017.

[53] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[54] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, and J. T. Barron, "Unprocessing images for learned raw denoising," in *CVPR*, 2019.

[55] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard/

[56] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "Hdr image reconstruction from a single exposure using deep cnns," *ACM transactions on graphics (TOG)*, vol. 36, no. 6, pp. 1–15, 2017.

[57] L. Meylan and S. Susstrunk, "High dynamic range image rendering with a retinex-based adaptive filter," *IEEE Transactions on image processing*, vol. 15, no. 9, pp. 2820–2830, 2006.