PERCEPTION FOR AUTONOMOUS DRIVING SYSTEMS UNDER LOW-LIGHT CONDITIONS

by

Muhammad Omer Farooq Bhatti

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Engineering in
Data Science and Artificial Intelligence

Examination Committee: Prof. Matthew N. Dailey (Chairperson)
Dr. Mongkol Ekpanyapong
Dr. Chaklam Silpasuwanchai

Nationality: Pakistani
Previous Degree: Master of Computer Science
Virtual University of Pakistan
Pakistan

Scholarship Donor: AIT Scholarships

Asian Institute of Technology
School of Engineering and Technology
Thailand
May 2023

AUTHOR'S DECLARATION

I, M. Omer Farooq Bhatti, declare that the research work carried out for this thesis was in accordance with the regulations of the Asian Institute of Technology. The work presented in it is my own and has been generated by me as the result of my own original research, and if external sources were used, such sources have been cited. It is original and has not been submitted to any other institution to obtain another degree or qualification. This is a true copy of the thesis including final revisions.

Date: April 18, 2023

Name: M. Omer Farooq Bhatti

Signature: 

ACKNOWLEDGEMENTS

I would like to thank Professor Matthew Dailey for his advice and constant support and guidance in the process of writing this thesis. I further thank Professor Mongkol Ekpanyapong and Professor Chaklam Silpasuwanchai for taking the time to be part of my thesis committee and providing valuable feedback.

I would also like to acknowledge all the people whose work this thesis builds on. Their work has been cited throughout this document and I appreciate their contributions to the field.

I thank the Asian Institute of Technology for the scholarship provided to me, which enabled me to carry out this work. Finally, I appreciate the love and patience of my family for supporting me and enabling me to pursue higher education.

ABSTRACT

A great deal of progress has been made recently on autonomous driving systems (ADS). However, despite this progress, we are as of yet unable to achieve fully autonomous driving capability without human supervision. The problem areas occur in the perception system of the ADS, where unexpected conditions can lead to a failure of the perception module, causing accidents. Such accidents undermine public trust in the technology and set back the adoption of fully autonomous vehicles. Many of the errors in perception are caused by variations on adverse lighting conditions. In this thesis, I explore the particular issue of low-light situations, in which, due to low exposure, normal methods for high-level vision tasks may not work as well. This thesis presents a comparative study of various state-of-the-art image enhancement models for the purpose of facilitating high-level vision tasks. The results enable us to evaluate the suitability of particular models for the required task under low-light conditions. In addition, joint-training of EnlightenGAN with YOLOv5 is performed. The resulting model enables us to learn image features which are more robust under varying illumination conditions as well as produce state-of-the-art results (79.5% mAP) for object detection under low-light conditions.

CONTENTS

LIST OF TABLES

Tables **Page**

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Background

Despite a great deal of progress made on autonomous driving systems (ADS), fully automated driving without the need for a human fail-safe still eludes the researchers in the field. A significant portion of the problem stems from issues in the system's perception modules such as those arising from low and varying illumination across images. Modern ADS's utilize multiple sensor modules to overcome this problem. However, this results in increased cost in R&D and manufacturing. Major ADS vendors such as Tesla have thus recently announced switches to camera only perception (Forbes, 2021).

Although ADS research has come very far since its inception, there are still unsolved problems left to be tackled. Real-time object detection (Jocher et al., 2022) and segmentation (L.-C. Chen, Zhu, Papandreou, Schroff, & Adam, 2018) are now practical, with very good accuracy under standardized conditions. But the benchmark datasets (Everingham, Gool, Williams, Winn, & Zisserman, 2010; Lin et al., 2014) on which these models are typically evaluated only contain good quality images of well-lit scenes. When the ADS encounters non-standard conditions, there is still significant potential for system failure as the model encounters out-of-domain data. Such conditions include disparity between training data and encountered conditions, low-light scenarios, illumination variance, glare/occlusion, and adverse weather conditions such as rain and fog.

Another issue is that most commonly used datasets for training object detection and segmentation models are gathered from Europe (KITTI) or the US (Yu et al., 2020). It is expected that under different driving environments, models trained on these datasets may not generalize well. Given this, it becomes critical to test these models on data from the local environment then further fine-tune the model if necessary, to ensure peak accuracy.

Currently, most open-source datasets addressing driving tasks do not have enough data for low-illumination situations. The Berkeley Deep Drive Dataset Yu et al. (2020) is currently the driving-related dataset with the most diverse driving conditions, including some low-illumination situations. The night driving scenes from this dataset can be used

to evaluate driving perception models in terms of their low-light performance.

In addition to this, most of the models currently being used for low-light enhancement need paired image data for training. Unfortunately, there is currently no driving-related dataset in the public domain that provides paired good-light/low-light images. Most of the research in this domain therefore utilizes synthetic low-light images, e.g., those created using the work of Liu, Breuel, and Kautz (2017). Very little research on this topic using real-world data has been done due to the difficulty of obtaining such paired data.

Within the area of low-light perception, object detection under low illumination conditions is a critical problem that needs to be addressed. Low light situations lead to a lack of information in the input image, which makes it harder for downstream models to learn patterns that will generalize well to unseen data (Loh & Chan, 2019). Previous approaches to this problem have utilized simple pre-processing methods such as histogram equalization, other image enhancement techniques, or raw sensor data to account for the low illumination conditions. There have also been approaches that explore end-to-end training (H. Guo, Lu, & Wu, 2021) and single-step approaches (Cui et al., 2021). The best current benchmark for object detection under low-light conditions is the ExDark (Loh & Chan, 2019) dataset.

Beyond the simple pre-processing methods mentioned above, another approach for low-light object detection is to use a more sophisticated image enhancement model on the low-light image prior to carrying out the high-level vision task. This has been demonstrated by C. G. Guo et al. (2020) for the face detection task. The reasoning behind the aim of enhancement as a pre-processing step is that low-light images may contain features that are not easily learned by a detection model such as YOLOv5, suggesting a pre-processing step that makes the features/information more discernible to the downstream model. However, these types of image enhancement methods are limited by the fact that they cannot generate information that is not already present in the input image; they can only enhance certain features to make them easier to learn by a model with specific capacity and a particular optimization objective. A single model, using domain adaption or a multi-tasking approach, could in principle enable greater knowledge sharing, which would result in better latent features, reducing the required number of overall model parameters and enabling faster processing (Sasagawa & Nagahara, 2020).

Related to this issue, the evaluation metrics currently used to benchmark image enhancement models, namely Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), are not reflective of the performance of the image enhancement models specifically for high-level vision tasks such as object detection or image segmentation. Training a separate model using task-agnostic criteria is not necessarily optimal for producing a processing stream that is proficient at high-level vision tasks. Evaluation protocols that consider both image enhancement and object detection together are needed.

1.2 Problem Statement

Current research in autonomous driving systems suffers from the limitations in perception experienced under low-light driving conditions. Also there is a lack of domain-specific dataset for driving, in order to enable supervised training of image enhancement models. There is also very little research into end-to-end training for object detection and segmentation in low-light images. Additionally, the evaluation metrics being used right now are unrelated to gauging the suitability of the image enhancement models for facilitating high-vision tasks.

1.3 Research Questions

The pertinent research questions this study explores are as following.

1. Which image enhancement method, from among the current state-of-the-art, performs best on the criteria of speed/accuracy for downstream high-vision tasks related to ADS?
2. How well do domain-specific light enhancement models, such as MBLLEN and KinD, perform in comparison to models with a self-supervised training approach for the driving-related vision tasks?

1.4 Objectives

The purpose of this work is to push the state-of-the-art on situation awareness for ADS in low-light conditions. The steps taken in order to accomplish this task are as follows.

1. Evaluate state-of-the-art low-light image enhancement models for their performance in a vision pipeline for autonomous driving tasks using publicly available datasets.
2. Explore joint training for object detection in low-light images using a low-light image enhancement model with a YOLOv5 model.

1.5 Limitations and Scope

There are no datasets that specifically provide driving-related images with low-light and corresponding normal-light versions. Due to this, currently available datasets that are not domain-specific will be used to train the supervised low-light image enhancement models for evaluation.

In addition, due to limitations of time, the scope of this study is limited to object detection and image segmentation tasks.

CHAPTER 2
LITERATURE REVIEW

There are generally two types of approaches taken for the design of autonomous driving systems as described in Yurtsever, Lambert, Carballo, and Takeda (2020). The first approach is a modular approach where the various components of autonomous driving systems are designed and trained individually and then integrated to deliver a complete solution. The second approach is an end-to-end driving system where low-level perception inputs are processed in a black box-like system and appropriate controls are output, for example via reinforcement learning.

The advantages of the modular approach are that it is more interpretable and allows for greater customization; in addition, engineers can work on different modules in parallel. The disadvantage of this approach is that increased human involvement in devising interim features for the system may not be optimal and may not lead to the best features for accomplishing the end-task. On the other hand, end-to-end driving systems are not interpretable. Such systems are based on learning a policy through experience, using techniques such as reinforcement learning, that is expensive to manage for an autonomous driving system. The system operates like a black box, where intermediate features and inputs from different sensors are combined without any human input.

For the purpose of this study, we assume a modular system approach, with an explicit perception module that provides high-level outputs for the control system.

2.1 Sensors used in Autonomous Driving

Modern Autonomous Driving Perception technology utilizes mainly three categories of sensors: Cameras, LiDARS, and RADARS. There are many types of cameras which are used for sensing the environment around the car. Most commonly, monocular or stereo cameras are used for autonomous driving applications. In this vein, the standard dataset used to benchmark autonomous driving research is the KITTI dataset (Geiger, Lenz, Stiller, & Urtasun, 2013). The KITTI dataset contains videos taken from stereo cameras, along with LiDAR data, and contains annotations for object detection, instance segmentation, visual odometry, and depth estimation tasks, among others. However, RGB-D cameras, omni-directional cameras and event cameras are also increasingly being used

Figure 2.1

Comparison of different sensing modalities.

Modality	Affected by Illumination	Affected by weather	Color	Depth	Range	Accuracy	Size	Cost
Lidar	-	✓	-	✓	medium (< 200m)	high	large*	high*
Radar	-	-	-	✓	high	medium	small	medium
Ultrasonic	-	-	-	✓	short	low	small	low
Camera	✓	✓	✓	-	-	-	smallest	lowest
Stereo Camera	✓	✓	✓	✓	medium (< 100m)	low	medium	low
Flash Camera [77]	✓	✓	✓	✓	medium (< 100m)	low	medium	low
Event Camera [78]	limited	✓	-	-	-	-	smallest	low
Thermal Camera [79], [80]	-	✓	-	-	-	-	smallest	low

Note. Reprinted from Yurtsever et al. (2020).

nowadays, and corresponding datasets are available to facilitate further research.

LiDARS are used in a sensor fusion technology, along with cameras and RADARs to provide a fool-proof mechanism for situation awareness. Camera sensors, while being full of information and cheaply available, have a number of failure scenarios. These involve situations where there is occlusion, illumination variation, low light areas, glare, snow, fog, rain and other such adverse weather conditions. In order to account for scenarios such as these, it is critical for safety of the vehicle to provide an additional sensing mechanism to augment the perception of the car.

LiDARs are perfect as redundancy for a perception module. LiDARs emit infrared lasers to gauge the distance of objects from the car. LiDARs work well at short distances; however, they are also unreliable under certain weather conditions. To provide further redundancy for such scenarios, RADARs can also be used in conjunction with the cameras and LiDARS. RADARs can operate over long distances and are generally unaffected by weather conditions. However, RADARs as a sensing modality do not have a high accuracy, especially compared to LiDARS.

The classic approach to optimally utilize the input from all the sensors (cameras, LiDARS and RADARS) is to use Kalman filters. However, lately, deep learning approaches such as FuseMod (Rashed et al., 2019) have also been tried with success.

2.2 Camera-based perception modules for Autonomous Driving System

Since the use of additional sensors and equipment is very expensive and drives up the cost of research and development of autonomous driving systems, a critical area of research in autonomous driving systems is using only camera-based image input to gather

perceptual information about the environment. The perception module of the ADS has to perform a number of different tasks such as road and lane detection, vehicle & pedestrian detection, and estimation of distance of detected objects from the ego vehicle. For the purpose of depth estimation, stereo cameras are used to obtain a pair of images of the same scene, offset by some small measure. Three-dimensional reconstruction is performed by matching keypoints/features in the two images to recover the corresponding pair of points and using them to reconstruct the 3D scene. The reconstruction is made possible by detecting features to match corresponding points or by using deep learning approaches to match the stereo images.

Some popular approaches to these driving-perception tasks are listed below.

1. Road & Lane Detection

   The first task for an ADS is to determine the driveable area for the vehicle and lane detection for navigating the traffic. There are multiple learning and non-learning approaches to accomplish this task. The relevant datasets for lane detection are CULane (Xingang Pan & Tang, 2018) and Berkeley Deep Drive Dataset (Yu et al., 2020).

2. Object Detection

   Object detection is required to detect the vehicles, pedestrians, traffic signals and any other objects in the pathway of the vehicle. Three-dimensional object detection is the process of detecting objects in the 3D point cloud space, using data gathered from the LiDAR, and projecting it in the 2D image space if needed. Two-dimensional object detection on the other hand is performed solely on the images captured from the camera. A popular deep learning model for 2D object detection is YOLOv5 (Jocher et al., 2022).

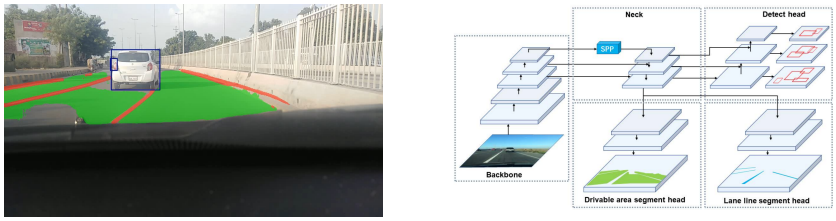
3. Image Segmentation

   Semantic segmentation provides a class definition for every pixel in the image. It is used to segment the image into regions at pixel-level. Instance segmentation provides a class and object definition for every pixel in the image. Segformer (Xie et al., 2021) and DeepLabv3-plus (L.-C. Chen et al., 2018) are popular semantic segmentation models, whereas Mask-RCNN (He, Gkioxari, Dollár, & Girshick,

Figure 2.2

YOLOP: A Panoptic Driving Perception Model

(a) *Sample output of YOLOP* **(b)** *YOLOP model architecture*



Note. Model architecture reprinted from Wu, Liao, Zhang, and Wang (2021).

2017) and YOLACT (Bolya, Zhou, Xiao, & Lee, 2019) are instance segmentation models.

4. Depth Estimation

   Images taken from stereo cameras are used to determine the disparity and reconstruct the 3D image space. Classical as well as deep learning techniques exist for this purpose. Classical techniques involve matching corresponding keypoints from the two images. Stereo matching neural network models such as PSMNet (Chang & Chen, 2018) have also been used successfully for the same purpose.

A multi-task approach to driving perception tasks have been made possible with the recent availability of large datasets such as BDD-100K (Yu et al., 2020), which provide a diverse driving dataset with annotations for many driving-related tasks such as object detection, image segmentation, driving area segmentation and lane detection. Provision of such datasets has enabled research in multi-task models for driving perception. A multi-task approach is beneficial in that it reduces the compute time and increases accuracy by sharing information for all the tasks in the parameters of the network. This results in less parameters for the network, due to knowledge sharing, and reduces the computation requirements.

Wu et al. (2021) present a unified model for panoptic driving perception using multi-task approach, called YOLOP. YOLOP utilizes a shared backbone and neck for feature extraction and aggregation, plus three different heads for individual tasks. YOLOP is used for object detection, driveable area segmentation and lane detection. As per the results presented by the author, it gives state-of-the-art performance for all three tasks, even

surpassing many task-focused models. Additionally, it was shown that YOLOP is much faster for inference, since it is able to operate at 41 FPS, which beats all other models in terms of efficiency. Another recent work exploring multi-task models is HybridNets (Vu, Ngo, & Phan, 2022), which while having more network parameters than YOLOP requires fewer computations and produces better accuracy for all three tasks.

2.3 Low-Light Image Enhancement

Research to overcome the above-mentioned limitations is ongoing. There have been several attempts to address the low-light image problem. Several learning methods have been proposed, including approaches based on an auto-encoder architecture, CNN-based models, Retinex Theory-based models, and generative models, among others.

2.3.1 Auto-encoder and CNN based methods

LLNet (Lore, Akintayo, & Sarkar, 2017) was one of the first deep learning approaches taken to the problem of illumination enhancement of images. It utilizes a stacked sparse deep Auto-encoder (SSDA) with three encoder layers and three decoder layers for image enhancement and de-noising. This approach is however quite computationally heavy, as the auto-encoder and decoder layers require many mathematical operations; furthermore, LLNet scores quite low on SSIM metric. MBLLEN (Lv, Lu, Wu, & Lim, 2018) is a CNN based model that optimizes the model on three loss functions: Structure loss, Context loss, and Region loss. MBLLEN gives state-of-the-art results for image enhancement; however, the model complexity is comparatively high, which disqualifies it for real-time usage.

2.3.2 Retinex Theory based methods

There has been quite a bit of research, including LightenNet (Li, Guo, Porikli, & Pang, 2018), RetinexNet (Chen Wei, 2018), and KinD (Zhang, Zhang, & Guo, 2019), that has explored the use of Retinex Theory for a deep learning solution to this problem. This approach seeks to decompose the image into reflectance and illumination maps, after which the reflectance map is passed through a de-noising mechanism and the illumination map is passed through an illumination-enhancing mechanism. The theory operates on the principle that the reflectance map of an image remains constant under any illumination conditions. Thus by image decomposition, a change in the illumination map can be made to obtain the enhanced image. Among these models, KinD is the best performing model both in terms of evaluation metrics and efficiency.

2.3.3 Other approaches

Other approaches include the use of adversarial training, deep curve estimation, and transformers to generate the enhanced image from a low-light image. Among these, EnlightenGAN (Jiang et al., 2021) and Zero-DCE (C. G. Guo et al., 2020) are self-supervised methods. Zero-DCE learns a mapping from low-light to enhanced images by estimating a curve. The parameters of the curve are only dependent on the input image and are optimized using four non-reference functions. This approach does not require a paired image training dataset, and thus it is much easier to adapt to different scenarios. EnlightenGAN utilizes adversarial training with a self-regularized perceptual loss function and a global-local discriminator to generate enhanced images. IAT (Cui et al., 2022) uses attention-based methods to convert the RGB image into raw data and tweak the parameters of ISP pipeline to reconstruct the enhanced image. This method only requires 90k parameters and has much smaller computation time. It also provides state-of-the-art results on LOL dataset. However, this approach does require a paired image dataset for training.

MAXIM (Tu et al., 2022) uses a MLP-based architecture to perform low-level vision tasks. It has produced state-of-the-art results for various vision tasks such as de-blurring, de-hazing, de-raining and image enhancement. However, MAXIM, being based on a MLP architecture, requires a large amount of computation and thus may not be suitable for real-time use. Ma, Ma, Liu, Fan, and Luo (2022) present SCI, a model that uses a self-calibrated module and unsupervised training loss in order to constrain the illumination learning and aid convergence between results of each stage in a multi-stage training operation. This model is much less complex in comparison to MAXIM and easily able to run in real time.

Most of these approaches require a dataset consisting of paired low-light and normal-light images, which can be hard to acquire. Datasets such as LOL (Chen Wei, 2018), MIT-Adobe FiveK (Bychkovsky, Paris, Chan, & Durand, 2011), and SID (C. Chen, Chen, Xu, & Koltun, 2018) are usually used for training this type of model. These datasets contain a large number of images of the same scene, captured using high and low exposure times. The images may be taken both indoors as well as outdoors and contain common objects and scenes, but lack diversity in terms of scenes.

Table 2.1

Comparison result of some low-light enhancement models on LOL dataset.

Sr.	Model	PSNR	SSIM
1	LLNet	17.959	0.713
2	LightenNet	10.301	0.402
3	RetinexNet	16.774	0.462
4	EnlightenGAN	17.483	0.677
5	MBLLEN	17.902	0.715
6	KinD	17.648	0.779
7	Zero-DCE	14.861	0.589

Note. Reprinted from Li et al. (2021).

2.4 Low-light Image Enhancement for High-vision tasks

There are at least two ways to address the problem that low-light and low-exposure images cause for high-level vision tasks. The first approach seeks to first improve the exposure of an input image so that it can then be fed through a deep neural network for another vision task on high-quality images. The second approach is to utilize a unified neural network to perform the high-level vision task (such as object detection) directly on low-light images.

Most current work takes the first approach. Improving the exposure of low-light images is potentially beneficial in many applications, not only for ADSs. Many researchers thus approach the problem as a general low-level vision task. However, these efforts have the downside of being designed primarily to perform well on image enhancement metrics and have not been integrated into high-level vision systems. Al Sobbahi and Tekli (2022) present a comparative study of some image enhancement and object detection models for low-light object detection, with YOLOv3 providing the best results.

Sasagawa and Nagahara (2020) explore the possibility of knowledge distillation and domain adaptation for training a unified YOLO model to “see in the dark." They utilize a SID model (C. Chen et al., 2018) fused with a YOLO backbone using glue layers in order to enable a YOLO model pre-trained on the MS-COCO dataset to detect objects in dark images. Since the SID model uses RAW image data as input, the composed model

likewise only utilizes RAW image data for detection. H. Guo et al. (2021) combine multiple image enhancement networks with Faster-RCNN (Ren, He, Girshick, & Sun, 2015) in order to improve low-light object detection.

Another approach has been taken by MAET (Cui et al., 2021), a multi-task model providing object detection in low-light images. To the best of our knowledge, this model provides the previous state-of-the-art results (mAP_{50} 74%) on the ExDark (Loh & Chan, 2019) dataset.

This thesis performs a comparative evaluation of some recent image enhancement models for low-light object detection using YOLOv5 detectors, and presents a new model obtained through joint-training that achieves state-of-the-art results.

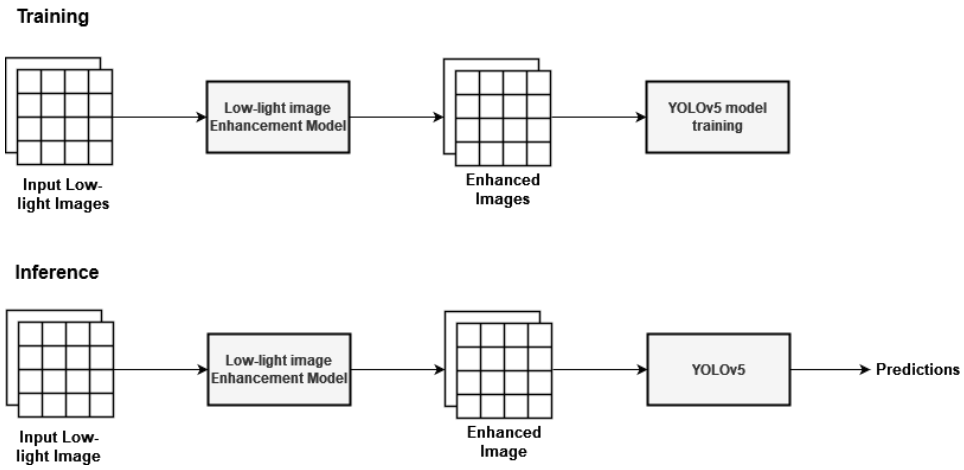
CHAPTER 3
METHODOLOGY

3.1 System Overview

In order to evaluate the performance of low-light enhancement models, we need to integrate the model into a pipeline for the relevant high-vision task. The relevant high-vision tasks for driving perception are object detection, image segmentation and lane detection. The current state-of-the-art models for these tasks are YOLOv5 (Jocher et al., 2022), YOLOv7 (Wang, Bochkovskiy, & Liao, 2022) for object detection; and Segformer (Xie et al., 2021), DeepLabv3-plus (L.-C. Chen et al., 2018) for image segmentation. We need to first enhance the input image by propagating it through a low-light enhancement model, and then feed it to the relevant task model. We follow this approach in this work. An alternative approach is to fuse the two models by utilizing features extracted through a low-light enhancement model in the backbone of the task model, as explored successfully by Sasagawa and Nagahara (2020).

Figure 3.1

Training and Inference Pipeline for experimentation.



3.2 System Design

There are two objectives for this study. The first objective is to perform a comparison of various image enhancement methods for downstream high-level vision tasks. The second objective is to jointly train a model to perform low-light object detection. The design for the two components is presented below.

3.2.1 Evaluation of image enhancement models

To evaluate the models, we need a labelled dataset of low-light images to serve as benchmark for the relevant tasks. Previous work in this domain by Sakaridis, Dai, and Van Gool (2021) and Loh and Chan (2019) presents us with two relevant datasets, ACDC (Adverse Conditions Dataset with Correspondences) and EXDark. EXDark is an object detection dataset compiled specifically as a benchmark for object detection in low-light images. ACDC is a benchmark for semantic segmentation of images under adverse weather conditions. It contains a large dataset of labelled images under many adverse conditions, including low-light conditions.

For comparison, we first need the baseline of how the object detection model, pre-trained on MS-COCO, performs on the EXDark dataset. Then we can examine the efficacy of fine-tuning by training the YOLO model on the training set from EXDark dataset. Finally, we utilize the low-light enhancement models to pre-process the images and using them for training the YOLO model; and then compare the test results. A similar approach is followed for the semantic segmentation model.

The low-light enhancement models to be used lie in two categories: supervised and unsupervised. The models trained using a supervised approach require a labelled dataset for training the model. The LOL dataset is predominantly used for this purpose. Most of the low-light image enhancement models use the LOL dataset for training and testing the models. LOL dataset is mostly composed of paired indoor images with low and normal light conditions. However, the images gathered in the ExDARK dataset are a mix of indoor and outdoor images.

Utilizing a supervised model presents us with the possibility that the model will not be generalizable to our target dataset. In addition, the models trained using the unsupervised approach generally do not perform as well on the evaluation metrics (PSNR, SSIM) used commonly, but the same cannot be said for downstream tasks. Given this, we perform a comparison of various supervised/unsupervised models in terms of their efficacy for the object detection task on low-light images in the ExDARK dataset. We use the standard evaluation metrics for object detection, such as mAP, in order to gauge the model's performance.

Below I provide a list of commonly used terms and define their use in the rest of this

thesis.

Low Light Refers to the original ExDark dataset and any YOLO model trained on this dataset.

Zero-DCE Deep learning model that learns a pixel-wise mapping to brighten the input image. May refer to the image enhancement model itself, the ExDark dataset enhanced by the Zero-DCE model, or any YOLO model trained on this enhanced dataset.

EnlightenGAN Generative model based on a UNet-like architecture for low-light image enhancement. May refer to the image enhancement model itself, the ExDark dataset enhanced by the EnlightenGAN model, or any YOLO model trained on this enhanced dataset.

EnlightenGAN-JT (ours) Our method for robust low-light object detection. May refer to the EnlightenGAN model jointly trained with the YOLOv5 model, the ExDark dataset enhanced by this model, or any YOLO model trained on this enhanced dataset.

The list provided above is not exhaustive and other models/methods mentioned in this document follow the same convention.

Experiment Design

The experimental design for this study is based on two independent variables: model and task. The variable model represents the different low-light models which will be used to enhance the training images. Tentatively, there are six levels for this variable, which represent six models to be evaluated. The models used are split between supervised and unsupervised/self-supervised. Task defines the downstream task that the images are used to perform: object detection or image segmentation. The dependent variable for the object detection task is mAP and for segmentation it is mIoU.

3.2.2 Joint training of EnlightenGAN and YOLOv5

In order to jointly train the EnlightenGAN model with the YOLOv5 model we need to first understand the EnlightenGAN architecture and training parameters. The Enlight-

enGAN model is based on the UNet (Ronneberger, Fischer, & Brox, 2015) architecture. However this UNet architecture is modified by adding an attention mechanism to the different layers of the model. This attention mechanism plays a key role in identifying darker regions of the image to focus on and helps avoid overexposure issues. The input of EnlightenGAN model is a 256×256 normalized image as well as an attention map to help guide the model training and inference. Figure 3.2 presents a visualization of the image[1], its normalized version and the calculated attention map.

(a) **(b)** **(c)**

Figure 3.2

Use of attention mechanism in EnlightenGAN a) Original image b) Normalized image c) Attention map.

The style transfer for the input images takes place by using four loss functions. A list of the loss functions along with a brief description is as follows.

The first class of loss functions is generator-related loss. Global loss for the generator is defined as,

$$L_G^{global} = \mathbb{E}_{x_f \sim \mathbb{P}_{fake}}[(D_{R_a}(x_f, x_r) - 1)^2] + \mathbb{E}_{x_r \sim \mathbb{P}_{real}}[D_{R_a}(x_r, x_f)^2],$$

whereas the local loss for the same is

$$L_G^{local} = \mathbb{E}_{x_r \sim \mathbb{P}_{fakepatches}}[(D(x_f) - 1)^2].$$

The self-feature-preserving loss, which is also a generator-related loss function and is applied both globally and locally, is defined as

$$L_{SFP}(I^L) = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^L) - \phi_{i,j}(G(I^L)))^2.$$

The second class of loss functions is related to the discriminator required for adversarial training. The relativistic function for global discriminator loss is

$$L_D^{global} = \mathbb{E}_{x_r \sim \mathbb{P}_{real}}[(D_{R_a}(x_r, x_f) - 1)^2] + \mathbb{E}_{x_f \sim \mathbb{P}_{fake}}[D_{R_a}(x_f, x_r)^2],$$

[1]Original image taken from video at https://www.youtube.com/watch?v=wqctLW0Hb_0

and the function for local discriminator loss is

$$L_D^{local} = \mathbb{E}_{x_r \sim \mathbb{P}_{realpatches}}[(D(x_r) - 1)^2] + \mathbb{E}_{x_f \sim \mathbb{P}_{fakepatches}}[(D(x_f) - 0)^2].$$

The self-feature-preserving loss is used to ensure fidelity between the features of output image and the input image. This loss is applied on a set of feature embeddings extracted from the input and output images using a VGG16 model (Simonyan & Zisserman, 2014) and its aim is to constrain the distance between these two embeddings. By ensuring similarity between the feature embeddings from the two images, the model inherently preserves feature similarity.

Since the loss has a global and a local component, the loss is applied iteratively to the whole image as well as a set of patches from the image. These image patches are of size 64×64 and are used to calculate the local components of self-feature-preserving loss, local generator loss and local relativistic discriminator loss. By using image patches to calculate these losses, model ensures more fine-grained generation of output features.

The hypothesis of this experiment is that by jointly training the enlightenGAN model with a YOLOv5 model and backpropagating the weighted YOLOv5 losses through the enlightenGAN model, we may be able to improve upon the object detection task. The model is expected to learn from the data and by utilizing the additional YOLOv5 loss, it may learn to generate features which further facilitate object detection. In order to do that, we need to remove the use of self-feature-preserving loss since it ensures similarity of input and output features.

The training pipeline for the experiment is designed as per Figure 3.3. The enlightenGAN model is connected to the YOLOv5 model via three 1×1 fully-convolutional "glue" layers with a skip connection. The YOLO loss is backpropagated through the network to optimize the enlightenGAN model.

In order for the weighted YOLOv5 loss to be effective in guiding the training, we need the YOLOv5 model to be able to detect the objects in the images generated by the enlightenGAN model. If we were to utilize a YOLOv5 model with randomly initialized weights, it would not be able to provide any useful gradient for enlightenGAN training, which in turn would affect the generation of good quality images, thereby hindering the optimization of the YOLOv5 model as well.

To facilitate convergence of training, we utilize a YOLOv5 model fine-tuned on the data pre-processed by the pre-trained enlightenGAN model. Because of the “domain shift" observed in the results of Experiment 1.2 in the Results section of this document, the fine-tuned YOLOv5 model was used instead of one pre-trained on MS-COCO dataset. By using the fine-tuned YOLOv5 model, we ensure that the target pixel distributions align as closely as possible.

An important thing to mention with regards to the training pipeline is the data augmentation methods employed during the training of the two models. As mentioned earlier, the EnlightenGAN model randomly crops the input image into 256×256 size for processing, whereas for optimum detection performance we need images of at least 416×416. In addition to this, the EnlightenGAN model performs data augmentation by injecting noise and performing random horizontal flips. Training of EnlightenGAN also requires a learning rate schedule which is pre-set by the original authors. Data processing and augmentation pipeline for YOLOv5 training is distinct from this and YOLOv5 training requires particular training parameters. In order to account for the training parameters of both models, the training is performed in stages. Also the training parameters of EnlightenGAN are adjusted such that they do not affect the detection pipeline.

The EnlightenGAN-JT model is produced in steps by first pre-training YOLOv5 on MS-COCO then fine-tuning it on the EnlightenGAN-enhanced dataset. Next, EnlightenGAN is trained jointly with this YOLOv5 model by backpropagating the YOLO loss through the input pixels then through the EnlightenGAN model. For joint training, the enhancement and detection modules are joined using three convolutional layers with a skip connection. The EnlightenGAN model is then frozen during re-training of YOLOv5 from scratch on the enhanced images.

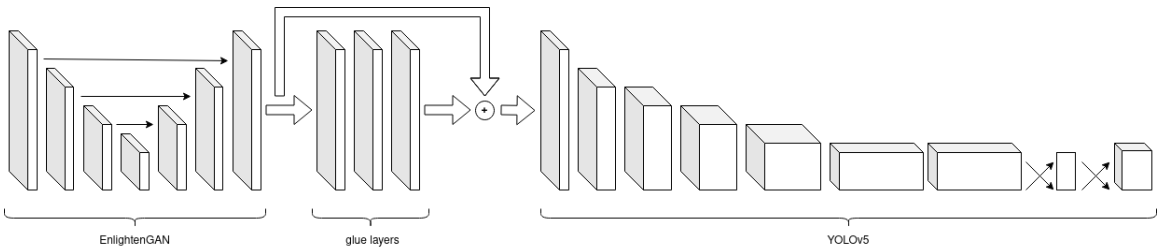


Figure 3.3

Diagram showing EnlightenGAN-JT training pipeline.

Figure 3.3 shows the architecture of the model during joint training, and Figure 3.4 shows

the training losses for the EGAN-JT generator model. “Real A" represents the low-light image from the ExDark dataset. “Real B" represents the image used for the style transfer of “Real A" image from low-light to brightness. “Fake B" is the resulting image generated by the model, with exposure corrected as per “Real B" and same features as “Real A". Discriminator losses are computed by comparing the “Real B" image with the “Fake B" image. Self-feature-preserving loss is calculated using L2 loss between the feature representations of the two images obtained using VGG16 model.

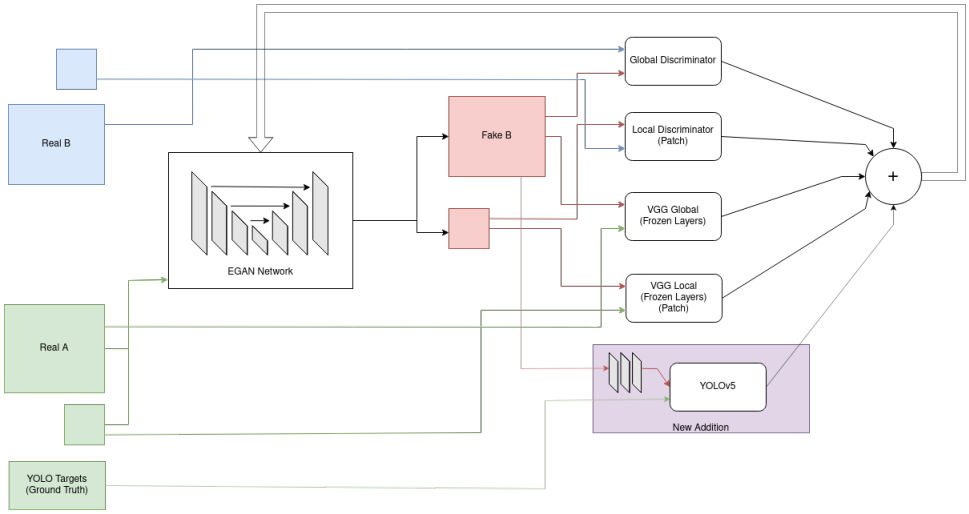


Figure 3.4

EnlightenGAN-JT training losses.

CHAPTER 4
RESULTS

This chapter presents the results of the aforementioned experiments. The first section presents the preliminary results, obtained by observing the results of image enhancement using various deep learning methods. Further to that, in Section 4.2, the first experiment explores the use of pre-trained YOLOv5 models. The second and third experiments, presented in Sections 4.3 and 4.4, explore the use of fine-tuned YOLOv5s and YOLOv5x models. The fourth experiment (Section 4.5) explores the robustness of the features learned by the fine-tuned YOLOv5x model. Section 4.8 presents an analysis of real-time performance by applying the image enhancement and detection pipeline on a video. Furthermore, the importance of the SFP (self-feature-preserving) loss in the training of EnlightenGAN is analyzed in Section 4.6. Section 4.7 presents results obtained by training YOLOv5s on a dataset enhanced by utilizing multiple low-light image enhancement models. The last two sections (4.9 & 4.10) present evaluation results of low-light enhancement for the semantic segmentation task. Detailed results with discussion for each experiment are given in the following sections.

4.1 Preliminary results

This section presents the enhancement results obtained using various deep learning methods.

4.1.1 KinD

Figure 4.1 shows an example of the enhanced output from KinD (pre-trained on the LOL dataset) model along with reflectance and illumination maps for low-light and enhanced images. As can be observed, the model does not generalize well from the LOL dataset to our example image.

4.1.2 MBLLEN

Figure 4.2 shows sample output of the MBLLEN model on an input low-light image from sample data. The middle image shows the output of the model, and the right image shows the image after contrast enhancement

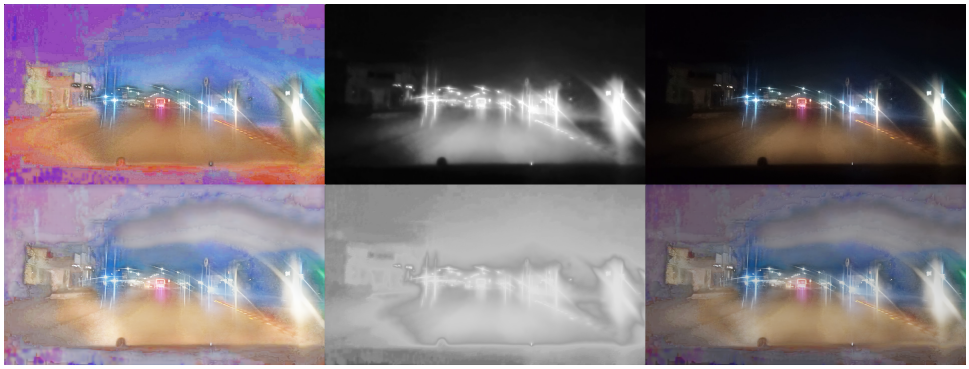


Figure 4.1

Output of the Kindling the Dark (KinD) model on an input image from our dataset.

Figure 4.2

Output of the MBLLEN model on an input image.

4.1.3 MAXIM

Figure 4.3 shows sample output of the MAXIM model (pre-trained on LOL dataset) on an input low-light image from our dataset.

(a) *Low-light image* **(b)** *Enhanced Image*

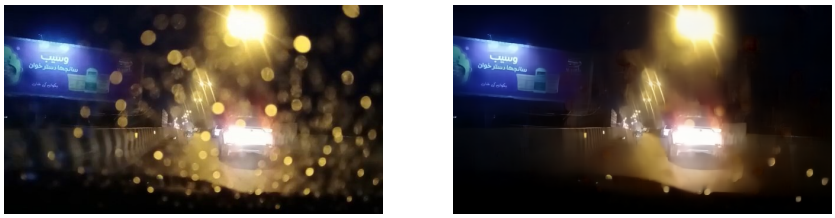
Figure 4.3

Image enhancement using the MAXIM model.

Figure 4.4 shows sample output of the MAXIM model on a rainy input image from our dataset.

4.1.4 EnlightenGAN

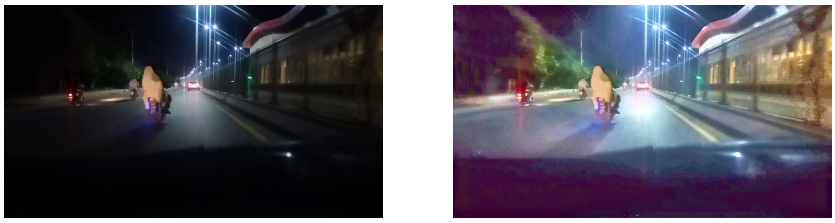
Figure 4.5 shows sample output of the EnlightenGAN model on an input low-light image.



(a) *Rainy image* **(b)** *De-rained Image*

Figure 4.4

Image de-raining using the MAXIM model.



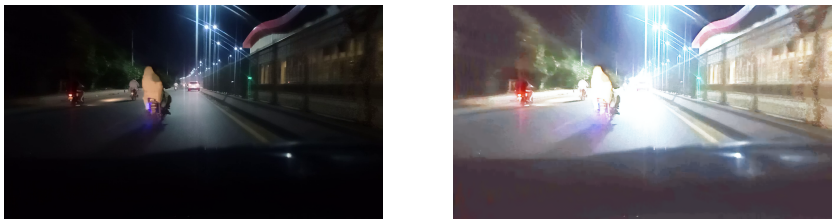
(a) *Low-light image* **(b)** *Enhanced Image*

Figure 4.5

Image enhancement using EnlightenGAN.

4.1.5 Histogram Equalization Priors

Figure 4.6 shows sample output of the HEP model on an input low-light image.



(a) *Low-light image* **(b)** *Enhanced Image*

Figure 4.6

Image enhancement using HEP.

4.1.6 Zero-DCE

Figure 4.7 shows sample output of the YOLOP model on a Zero-DCE enhanced video. The pre-trained Zero-DCE model was used for video enhancement without fine-tuning on domain-specific data.

(a) *Low-light image* **(b)** *Z-DCE enhanced Image*

Figure 4.7

Comparison of YOLOP detection in a Zero-DCE enhanced image.

4.2 Evaluation of pre-trained YOLOv5 models on the ExDark dataset

This section presents evaluation of pre-trained YOLOv5 models on the test set of the ExDark dataset. The YOLOv5 models were pretrained on the MS-COCO dataset. I evaluated the models on two versions of the dataset. The first is the original dataset, and the second dataset has been pre-processed using the EnlightenGAN model.

The ExDark dataset consists of 20 object classes. The train-val-test split of the dataset is given in Table 4.1. The images in the dataset are classified in ten categories based on their illumination profiles as shown in Table 4.2.

The MS-COCO dataset consists of 80 object categories, including all the classes from the ExDark dataset. In order to evaluate the pre-trained YOLOv5 models on the ExDark dataset, we first need to re-label the ExDark dataset using the class ids from the MS-COCO dataset.

Table 4.1

Training, validation and test split of the images in the ExDark dataset.

Dataset	Training	Validation	Test	Total
ExDark	3000	1800	2563	7363

Table 4.3 shows the results of the evaluation of pre-trained YOLOv5 models on the original ExDark test set.

Table 4.4 shows the results of the evaluation of pre-trained YOLOv5 models on the enhanced ExDark dataset. The ExDark dataset, in this case, has been enhanced using the pre-trained EnlightenGAN model prior to evaluation using YOLOv5 models.

Table 4.2

Illumination profiles of the images in the ExDark dataset.

Sr.	Illumination	Training	Validation	Test	Total
1	Strong	999	522	538	2059
2	Weak	298	175	232	705
3	Single	316	191	255	762
4	Object	235	119	144	498
5	Ambient	716	457	925	2098
6	Low	51	56	134	241
7	Twilight	147	95	174	416
8	Shadow	28	14	14	56
9	Window	141	103	119	363
10	Screen	69	68	28	165
11	All	3000	1800	2563	7363

The results from the tables above show that the pre-trained models perform slightly worse on the dataset pre-processed using EnlightenGAN. We can assume that despite enhancing the illumination in the images, the propagation of images through EnlightenGAN introduces a slight domain shift that comes from the change in distribution of pixel intensities in the images. Consequently, we see that the pre-trained YOLOv5 models have a drop of about 0.9% on the pre-processed dataset compared to the original.

Table 4.3

Experiment 1.1 results. Evaluation of pre-trained YOLOv5 models on the original ExDark dataset.

Class	YOLOv5s	YOLOv5l	YOLOv5x
Person	69.0	77.0	**77.3**
Bicycle	69.0	76.4	**78.2**
Car	66.3	77.8	**79.5**
Motorcycle	52.5	**66.7**	66.3
Bus	82.7	93.4	**94.1**
Boat	43.5	**57.7**	56.0
Cat	57.3	70.9	**72.2**
Dog	71.3	83.7	**85.9**
Bottle	60.5	67.2	**66.7**
Cup	53.9	64.6	**68.7**
Chair	50.8	59.9	**61.7**
Table	20.5	27.8	**28.2**
All	58.1	68.6	**69.6**

Note. Reported numbers are mAP_{50} on the ExDark test set.

Table 4.4

Experiment 1.2 results. Evaluation of pre-trained YOLOv5 models on ExDark dataset enhanced using EnlightenGAN.

Class	YOLOv5s	YOLOv5l	YOLOv5x
Person	67.1	76.5	**77.2**
Bicycle	71.1	77.4	**77.7**
Car	65.3	74.6	**77.7**
Motorcycle	49.2	67.1	**66.9**
Bus	82.6	**93.6**	92.4
Boat	46.9	**58.5**	56.5
Cat	54.6	69.0	**72.4**
Dog	67.5	80.6	**83.7**
Bottle	59.1	65.8	**66.0**
Cup	52.1	63.3	**66.8**
Chair	48.1	57.5	**58.4**
Table	22.5	27.3	**29.1**
All	57.2	67.6	**68.7**

Note. Reported numbers are mAP_{50}.

4.3 Evaluation of fine-tuned YOLOv5 models on the ExDark dataset

The baseline mAP_{50} for low light object detection using the pretrained YOLOv5 models on the ExDark dataset is 58.1 and 69.6 for YOLOv5s and YOLOv5x, respectively. Naively, we might assume that enhancing the input images using a low-light enhancement module, such as EnlightenGAN, will result in an improved mean average precision score. But, as observed in the results from Table 4.4, this introduces a domain shift in the images and reduces the resulting accuracy.

As followup, I fine-tuned the YOLOv5s model on the ExDark dataset using the default hyperparameters for the MS-COCO dataset. The dataset was pre-processed using various image enhancement models, and the original dataset serves as a baseline for comparison.

As can be observed in Table 4.5, the pre-processing of the images using an image en-

hancement model results in improved mean average precision results. The best image enhancement model in these settings comes out to be Zero-DCE, which is trained in a self-supervised manner. We obtain an improvement of 1.6% over the baseline (i.e without image enhancement). The second best image enhancement model, in terms of accuracy, is MBLLEN, which is trained using a paired image dataset (LOL) in a supervised manner. Thus we can see that there is no significant advantage in choosing a model trained under a self-supervised regime over one which is not. However, in terms of inference speed, the MBLLEN model is not as fast as Zero-DCE or EnlightenGAN. Additionally, the self-supervised training for EnlightenGAN and Zero-DCE frees us from the need for a domain-specific paired image dataset which enables their deployment in more use cases.

Table 4.5

Experiment 2 results. Evaluation of fine-tuned YOLOv5s model on enhanced ExDark dataset.

Class	Original	EnlightenGAN	Zero-DCE	MBLLEN	KinD	IAT
Person	66.4	65.8	67.1	**67.2**	**66.0**	67.0
Bicycle	**73.3**	73.9	**76.1**	74.9	73.9	73.8
Car	**67.9**	69.5	**70.6**	69.9	69.4	69.3
Motorcycle	50.7	**53.8**	49.6	51.7	**49.4**	51.7
Bus	80.3	81.5	82.7	**85.1**	82.2	**79.5**
Boat	**63.6**	64.4	**65.1**	**63.6**	64.9	**65.1**
Cat	56.7	55.6	**57.4**	**54.9**	55.8	57.1
Dog	67.8	**70.1**	68.9	67.2	**65.0**	69.3
Bottle	59.2	**63.7**	62.4	62.7	**57.1**	58.8
Cup	53.6	53.8	56.0	**56.8**	**49.9**	54.4
Chair	54.6	53.9	54.8	**56.8**	**52.1**	54.1
Table	**34.0**	35.9	**37.4**	36.3	34.2	35.6
All	60.7	61.8	**62.3**	62.2	**60.0**	61.3

Note. Blue indicates the best result among the different models, and red indicates the worst. Reported numbers are mAP_{50}.

In addition to YOLOv5s models, we also train the more complex YOLOv5x models on the enhanced datasets in order to observe the effect of increased number of parameters on

accuracy. Figure 4.8 shows examples of images processed by the three image enhancement models along with the original image from the ExDark dataset. From the images, it can be observed that the Zero-DCE and EnlightenGAN-JT images are more color consistent with the original image as compared to the EnlightenGAN. Also jointly training the EnlightenGAN with a YOLOv5 model (EnlightenGAN-JT) results in a model which outputs images with illumination levels considerably less than the original EnlightenGAN output. It can be inferred that during joint training, the EnlightenGAN-JT model only adjusts illumination levels up to the point required to minimize YOLOv5 losses.

Table 4.6

Experiment 3 results. Evaluation results for YOLOv5 models fine-tuned on the ExDark dataset.

Sr.	**Model**	YOLOv5s		YOLOv5x	
		mAP_{50}	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$
1	Low-light	66.1	35.4	78.1	45.1
2	EnlightenGAN	66.2	36.0	78.1	48.1
3	MBLLEN	63.9	34.6	79.3	47.7
4	Zero-DCE	65.9	35.7	79.4	**49.7**
5	EnlightenGAN-JT	67.8	**38.9**	**79.5**	47.3
6	IAT	66.6	37.0	77.4	47.3
7	SCI	**68.2**	38.6	78.9	45.8

Note. Results of evaluation on the ExDark test set. Models have been optimized for the mAP_{50} metric.

The YOLOv5x models were optimized on the criterion of best mAP_{50}. The results are presented in Table 4.6. The evaluations are presented in terms of two metrics: mAP_{50} and $mAP_{50:95}$. YOLOv5x trained on the EnlightenGAN-JT dataset has the highest mAP_{50} value, with Zero-DCE dataset being second best. However, YOLOv5x trained on the Zero-DCE enhanced dataset gives the best $mAP_{50:95}$ accuracy (an increase of 4.6%). In the case of mAP_{50}, we obtain an increase of 1.4% over the original (low light) dataset by utilizing the EnlightenGAN-JT dataset.

Figure 4.9 shows a confusion matrix for the EGAN-JT YOLOv5x model validation results. From the confusion matrix we can see that the model has high false negative (FN)

(a) *Low-light image*

(b) *Z-DCE enhanced image*

(c) *EGAN enhanced image*

(d) *EGAN-JT enhanced image*

(e) *MBLLEN enhanced image*

(f) *IAT enhanced image*

(g) *SCI enhanced image*

(h) *KinD enhanced image*

Figure 4.8

Examples of image enhancement using various low-light enhancement models.

rates for “Boat,” “Bottle,” and “Table” object classes and high false positive (FP) rate for “People” object class.

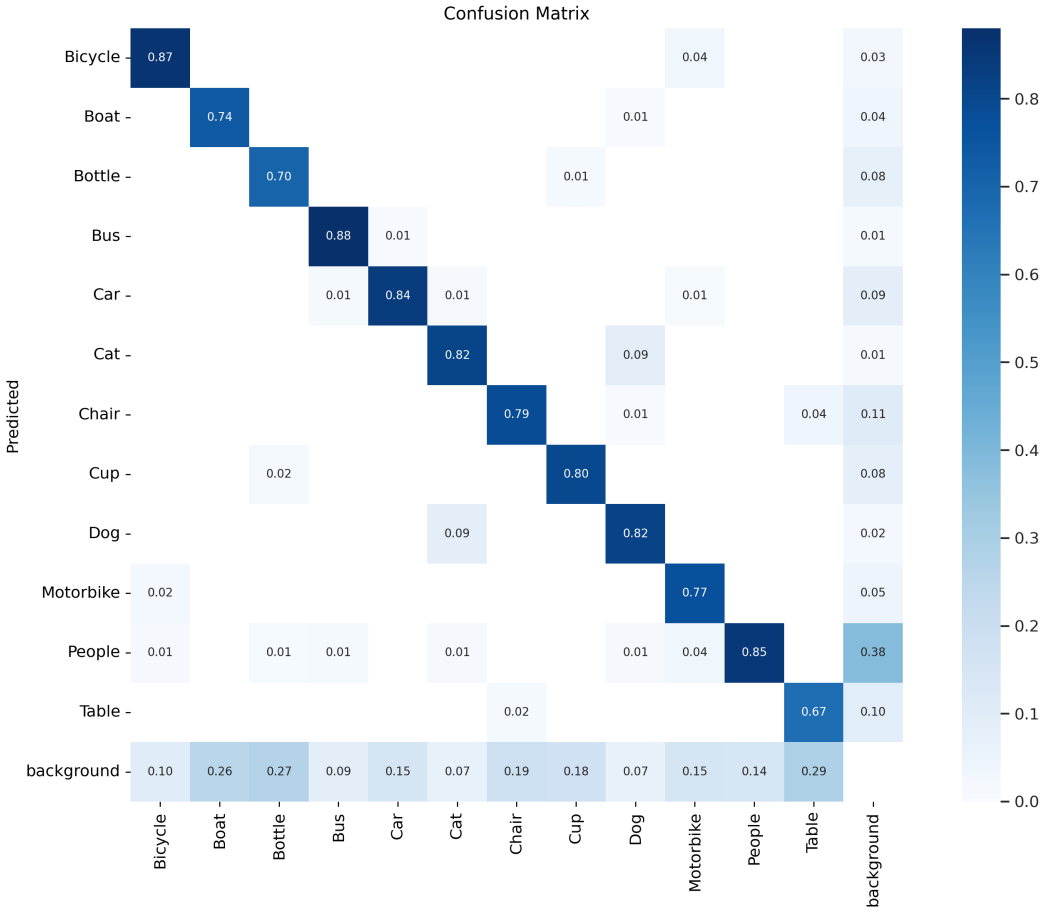


Figure 4.9

EnlightenGAN-JT YOLOv5x evaluation confusion matrix.

4.4 Evaluation of fine-tuned YOLOv5x models under simulated lighting conditions

The reasoning behind applying image enhancement as a pre-processing step is that the low-light images contain features that are not easily learned by the YOLOv5 model. I thus apply a pre-processing step in order to make the features/information more discernible to the downstream model. However, the image enhancement step is limited by the fact that it cannot generate the information that is not present in the input image; it can only enhance certain features to make them easier to learn with a limited model capacity and a certain optimization objective. Given this, questions arise as to how robust are the features learnt by the YOLOv5 models trained on these processed datasets, and how well do they generalize to different input pixel distributions.

In order to evaluate this, I tested all of the YOLOv5x models on the original ExDark dataset without applying any image enhancement. For further analysis, the evaluation results of all the models are broken down with respect to the illumination profiles of the images. The results are presented in Table 4.7.

EnlightenGAN-JT model is clearly able to generalize best in this scenario. Both Zero-DCE and EnlightenGAN post mAP_{50} numbers lower than the baseline (original low light). We see that the features learned by the YOLOv5 model from the EnlightenGAN-JT dataset are generalizable even if we do not apply the low-light image enhancement pre-processing step.

Although the Zero-DCE model performs nearly as well as EnlightenGAN-JT when the pre-processing step has been applied, the difference becomes wider when we remove that step. The implication of these results is that the YOLOv5 model trained on the Zero-DCE dataset learns features which are not as robust to different lighting conditions (input pixel distribution) as EnlightenGAN-JT.

Table 4.8 further examines the results of the EnlightenGAN-JT model on the original (low light) vs the EnlightenGAN-JT enhanced dataset. For all of the illumination profiles, the evaluation on the enhanced dataset shows that the EnlightenGAN-JT model performs comparatively better on this dataset, except for the “Screen” and “Shadow” lighting conditions.

I further tested the robustness of the models by simulating different illumination condi-

tions using gamma correction on the test dataset. The results are featured in Figure 4.10. The results show that across a range of gamma values, the EnlightenGAN-JT model tends to generalize better to both darker and brighter images. Neither Zero-DCE nor EnlightenGAN generalize well to the darker images (with gamma values less than one). The YOLOv5 model trained on the original dataset (low light) is more accurate on the very dark images but gives lower mAP for other gamma values.

We can also use model ensembling to improve upon the $mAP_{50:95}$ metric (Table 4.9), which implies that some of the models are more effective for a certain subset of images and that averaging over the four models gives us an improved result.

Table 4.7

Experiment 4.1 results. Evaluation of YOLOv5x models on original low-light ExDark dataset.

Sr.	Lighting	Low-light	EGAN-JT	ZDCE	SCI	EGAN
1	Strong	73.2	**74.1**	71.0	70.6	70.8
2	Ambient	**81.3**	81.0	78.8	80.1	79.8
3	Single	**84.8**	82.3	79.7	79.6	81.4
4	Weak	70.9	**73.2**	69.5	68.3	67.4
5	Object	72.8	75.7	76.2	**78.2**	74.4
6	Twilight	**85.8**	81.2	80.7	84.5	79.8
7	Window	**79.8**	79.3	77.9	68.3	**79.8**
8	Low	72.3	**75.3**	71.1	73.3	67.5
9	Screen	87.7	88.8	**89.4**	88.2	89.0
10	Shadow	75.9	**78.0**	68.0	77.0	75.4
11	All	78.1	**78.5**	75.8	76.3	76.0

Note. Reported numbers are mAP_{50} on the original ExDark test dataset.

Table 4.8

Experiment 4.2 results. Evaluation of EGAN-JT YOLOv5x model on ExDark dataset with and without enhancement.

Sr.	Lighting	Low-light dataset	EGAN-JT dataset
1	Strong	74.1	**75.4**
2	Ambient	81.0	**81.5**
3	Single	82.3	**84.0**
4	Weak	73.2	**75.3**
5	Object	75.7	**77.7**
6	Twilight	81.2	**81.4**
7	Window	79.3	**80.1**
8	Low	75.3	**76.9**
9	Screen	**88.8**	87.6
10	Shadow	**78.0**	70.0
11	All	78.5	**79.5**

Note. Reported numbers are mAP_{50}.

Model	$\mathbf{mAP}_{50:95}$
Low Light	45.1
EGAN-JT (mine)	47.7
ZDCE	47.5
EGAN	46.6
Ensembling	**49.2**

Table 4.9

Experiment 4.4 results. YOLOv5x model ensembling improves $mAP_{50:95}$.

4.5 Testing the generalizability of our jointly trained EnlightenGAN-JT YOLOv5x model

Loh and Chan (2019) conclude that the objects in images under low-light conditions present significantly different features to a deep learning model than those captured in well-lit images. The authors show a clear demarcation in feature space between learned representations from low-light and normal light images. Given this, the question arises as to how robust are the features learnt by YOLOv5 models trained on these processed datasets and how well they generalize to different lighting conditions.

In order to answer this question, I tested all the YOLOv5x models on the original ExDark dataset without applying any image enhancement (Experiment 4.1). For further analysis, the evaluation results of all the models are broken down with respect to the illumination profiles of the images. The results are presented in Table 4.7. The EnlightenGAN-JT model is clearly able to generalize best in this scenario. We can deduce that the features learned by the YOLOv5 model from the EnlightenGAN-JT dataset are generalizable even if we do not apply the low-light image enhancement pre-processing step.

In Experiment 4.2 (Table 4.8) I further examine the results of applying the EnlightenGAN-JT model to the original (low light) dataset and the EnlightenGAN-JT enhanced dataset. The evaluation on the enhanced dataset shows that the EnlightenGAN-JT model performs better, except for the "Screen" and "Shadow" lighting conditions.

I further test the robustness of the models by simulating different illumination conditions using gamma correction on the test dataset prior to detection (Experiment 4.3). Gamma

correction is a non-linear operation that adjusts illumination values. For $\gamma > 1.0$, brightness is increased, and conversely, for $\gamma < 1.0$, the image is made darker. The value of γ must be greater than zero. The equation for the non-linear gamma correction operation on an image is

$$I_{out} = (\frac{I_{in}}{255})^{\gamma} \times 255.$$

Results of gamma correction are featured in Figure 4.10. The results show that for a range of gamma values, the EnlightenGAN-JT model tends to generalize better to both darker and brighter images. The YOLOv5 model trained on the original (low light) dataset is more accurate only on the darkest images. Figure 4.11 shows the comparison between Low-light, EGAN, and EGAN-JT models in terms of robustness of learned features.

Finally, we can also use model ensembling to improve upon the $\text{mAP}_{50:95}$ metric (Table 4.9), which implies that some models are more effective for certain subsets of images than others; averaging over the four models gives us an improved result (Experiment 4.4).

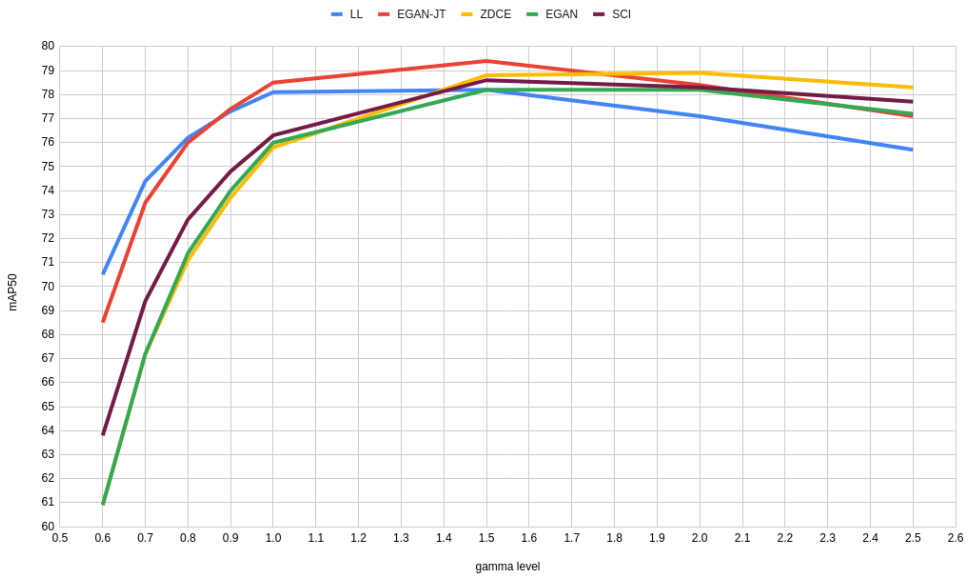


Figure 4.10

Experiment 4.3 results. Evaluation after gamma correction of images.

In addition to the mAP measurement, YOLOv5s features for a given input image (Figure 4.12) were also analysed. YOLOv5s models consist of 224 layers and 7,266,973 parameters. Feature vectors of size (512, 20, 15) were extracted from stage23_C3 of the

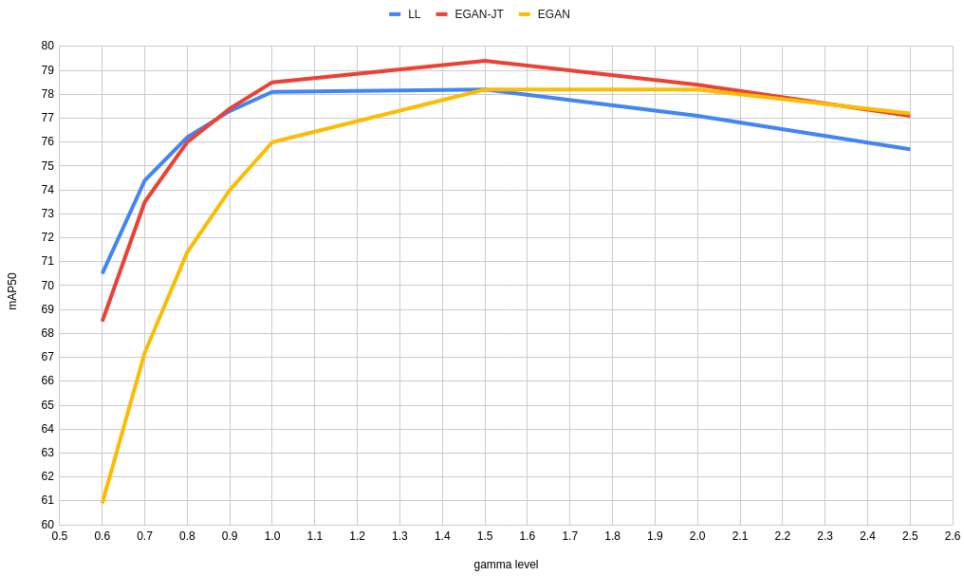


Figure 4.11

Experiment 4.3 results. Evaluation after gamma correction of images.

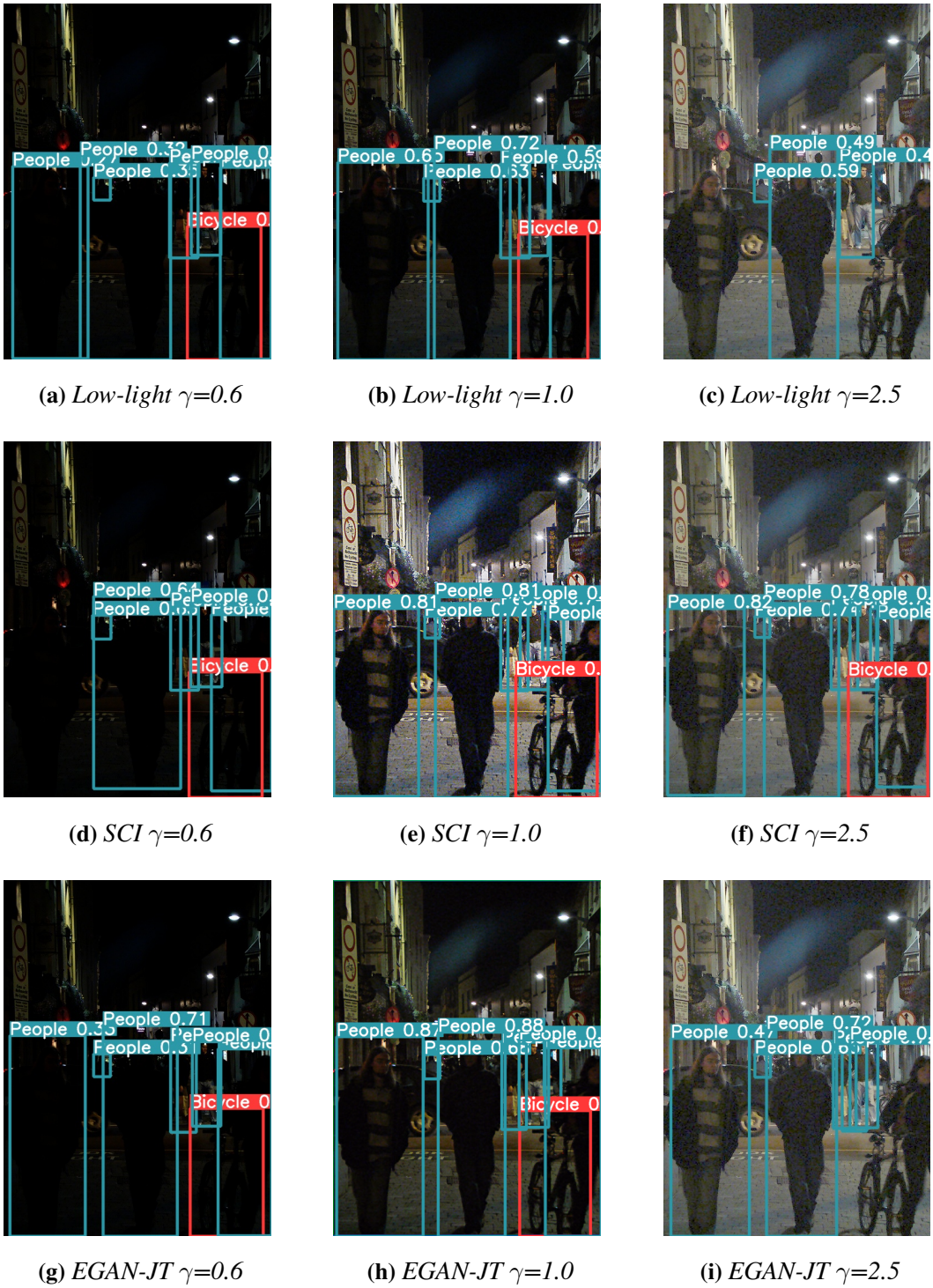
YOLOv5s model. In order to visualize the feature vectors on a 2D plot, a dimensionality reduction technique called principal component analysis (PCA) was utilized. PCA allows us to visualize high-dimensional data and describe variability in the observations more simply by projecting them in to a lower dimensional subspace.

Model	similarity($I_{\gamma=0.6}, I_{\gamma=1.0}$)	similarity($I_{\gamma=2.5}, I_{\gamma=1.0}$)	$\lvert\delta\rvert$
Low Light	0.896	0.818	0.078
ZDCE	0.799	0.956	0.157
EGAN	0.793	0.941	0.148
EGAN-JT	0.836	0.813	**0.023**
MBLLEN	0.845	0.915	0.07
IAT	0.798	0.935	0.137
SCI	0.827	0.950	0.123

Table 4.10

Cosine similarity between YOLOv5s features of images under different illumination conditions.

The resulting 2D points for the feature vectors of all the YOLOv5s models are plotted in Figure 4.13. Each plot in the figure visualizes features for a set of images and represents a detection pipeline. The dot represents the features of original images, the cross rep-



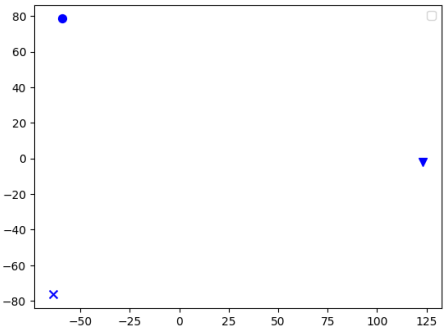
(a) *Low-light γ=0.6* **(b)** *Low-light γ=1.0* **(c)** *Low-light γ=2.5*

(d) *SCI γ=0.6* **(e)** *SCI γ=1.0* **(f)** *SCI γ=2.5*

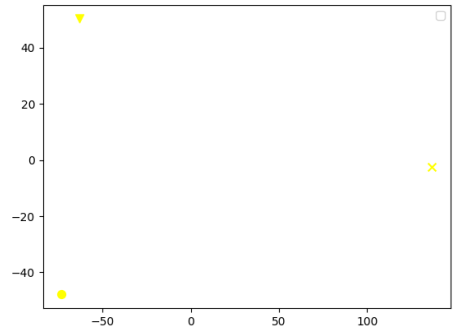
(g) *EGAN-JT γ=0.6* **(h)** *EGAN-JT γ=1.0* **(i)** *EGAN-JT γ=2.5*

Figure 4.12

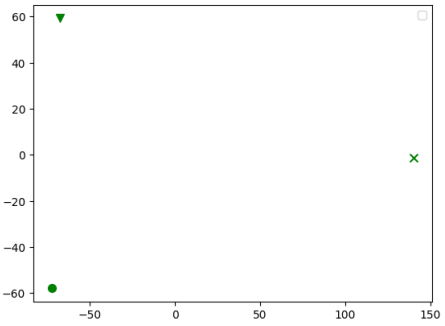
Detections using select low-light enhancement models.



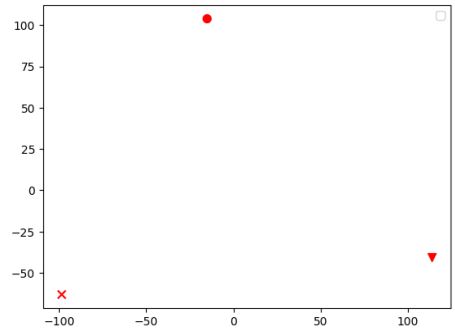
(a) *Low-light representations*



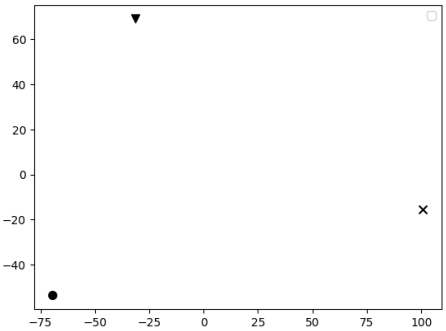
(b) *Z-DCE enhanced representations*



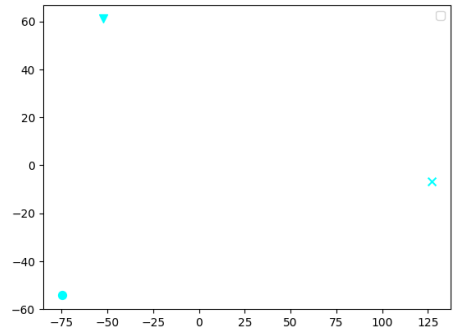
(c) *EGAN enhanced representations*



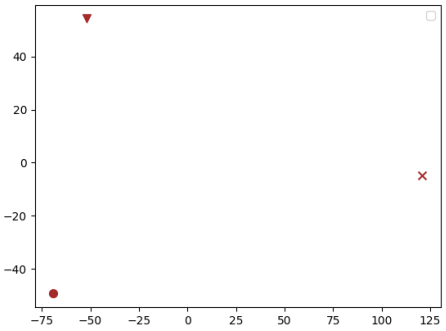
(d) *EGAN-JT enhanced representations*



(e) *MBLLEN enhanced representations*



(f) *IAT enhanced representations*



(g) *SCI enhanced representations*

Figure 4.13

Plots of image representations using various low-light enhancement models.

resents images with a γ correction of 0.6, and the triangle represents features of images with a γ correction of 2.5.

In a similar vein, in order to quantify the similarity of the representations, cosine similarity was used on the extracted feature representations for all the models. Cosine similarity is a measure of the similarity between two vectors and is defined by the below equation.

$$\cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{A}| \times |\mathbf{B}|}$$

The results are presented in Table 4.10. The cosine similarity is measured between the original image representation and the two gamma corrected versions of the images. From the table we see that the similarity of features between the darkened (γ=0.6) and the original image is highest for the "Low light" model. This is interesting as we also observed in the mAP evaluation that for γ=0.6, the "Low light" model performs best. We see that the second best similarity measure is for the MBLLEN model, which seems to imply that it should hold a similar position in the mAP evaluations as well, but this does not occur. The EGAN-JT model, which has the third-best similarity between features of these images, provides a better mAP at this gamma value than MBLLEN.

The third column in Table 4.10 presents the similarity between features of original and brightened (γ=2.5) images for all the detection models. The similarity of features is highest for ZDCE and SCI, which is consistent with the mAP results. However, we see that the similarity metric for the "Low light" model features is higher than those of EGAN-JT, which is contradictory to the mAP results.

Given this, we cannot say anything conclusive about the detection performance and cosine similarity between image features. However, it is to be noted that the difference in similarity of features, as represented by the fourth column in Table 4.10, is lowest for the EGAN-JT model. Nevertheless, based only on this analysis, it cannot be said conclusively that this is indicative of robustness of the detection model. The mAP results provide a much more solid foundation for that claim.

4.6 Training EnlightenGAN-JT with SFP loss

In the previous experiments, we utilized an EnlightenGAN-JT model trained without self-feature-preserving loss. In this section, for completeness, we conduct an experiment in which we train an EnlightenGAN-JT model with the self-feature-preserving (SFP) loss. Table 4.11 presents the validation results of the two models during the training process.

Model		**mAP**$_{50}$
EnlightenGAN-JT	with SFP loss	58.5
	without SFP loss	59.1

Table 4.11

Experiment 4.5 results. Validation results of EnlightenGAN-JT training.

Model		**mAP**$_{50}$	**mAP**$_{50:95}$
EnlightenGAN	pre-trained	66.2	36.0
EnlightenGAN-JT	with SFP loss	65.9	36.8
	without SFP loss	**67.8**	**38.9**

Table 4.12

Experiment 4.5 results. Evaluation results of EnlightenGAN-JT YOLOv5s on ExDark test set.

The trained EnlightenGAN-JT models were then taken and used to pre-process the dataset for the training of the YOLOv5s model. The comparison of the YOLOv5s model trained on these datasets is presented in Table 4.12.

From the results, we can see that the SFP loss adversely effects the mAP of the YOLOv5 model, but joint training without SFP significantly improves the results over using just the pre-trained model. This result suggests that the EGAN-JT model, without the SFP constraint, is able to adapt the statistics of the feature maps to obtain a better mAP.

Looking at the images (Figure 4.14) produced by the different variations of Enlighten-GAN, we can see a change in saturation levels in the output images. The EGAN-JT without SFP loss is closest to the original image. Introducing SFP loss in addition to the YOLO loss seems to drop the saturation levels a bit and raises illumination levels more.

(a) *Low-light image*

(b) *EnlightenGAN (pre-trained)*

(c) *EGAN-JT with SFP loss*

(d) *EGAN-JT without SFP loss*

Figure 4.14

Examples of image enhancement using EnlightenGAN-JT with and without SFP loss.

4.7 Training on multiple enhancements by different image enhancement models

As observed in the results of Experiment 4.4, model ensembling enables us to improve upon $mAP_{50:95}$ by utilizing models trained on several image enhancement methods. Considering this further, we might wonder whether training on a combined dataset enhanced by multiple image enhancement models will present us with further improved detection results. To this end, I chose two low-light image enhancement models, along with the original images, to test this thesis. The selection was made on the basis of results from Experiment 4.3, which showed that the original dataset works best for the darker images and that Zero-DCE enhancement yields the greatest accuracy for the brighter images. As the third set, I added the dataset enhanced with the EnlightenGAN-JT model which provides the best accuracy and robustness across a range of lighting conditions. New YOLOv5s and YOLOv5x models were trained on the images pre-processed by the two models as well as the original data. The results are presented in Table 4.13.

Pre-processing methodology	YOLOv5s		YOLOv5x	
	mAP_{50}	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$
Low-light (original)	66.1	35.4	78.1	45.1
Zero-DCE	65.9	35.7	79.4	49.7
EnlightenGAN-JT (mine)	67.8	38.9	**79.5**	47.3
All	**68.9**	**39.5**	79.3	**50.6**

Table 4.13

Experiment 4.6 results. YOLOv5 model trained on multiple image enhancement methods.

As can be seen in Table 4.13, training on multiple image enhancement models provides the best mAP_{50} of 68.9 (YOLOv5s), whereas among the single low-light enhancement models, SCI provided the best accuracy of 68.2%. These results suggest that training on multiple low-light image enhancement models' outputs enables learning of different sets of features that facilitate generalization of object detection. However, a possible caveat to this notion is the fact that since three types of pre-processed images were used to construct this dataset, it was composed of three times the size of the original. Thus, the increase in accuracy may not only be due to the image pre-processing methodologies, but also due to the effective increase in number of iterations the model was trained for in comparison to others. Furthermore, the same kind of increase in accuracy is not

observed for the larger YOLOv5x models. For YOLOv5x model, training on this combined dataset yields accuracy lower than the individual results of EnlightenGAN-JT and Zero-DCE, for the mAP_{50} metric.

4.8 Realtime Inference

Table 4.14 presents the average speed of inference for the proposed data processing pipeline in terms of frames per second. The vision pipeline was tested on a system with i7-6700 CPU, 16GB RAM and a 12GB RTX-3060 NVIDIA GPU. The pipeline involves image enhancement (except for low-light), detection, and tracking. The YOLOv5s model is used as the detector, and ByteTrack (Zhang, Wang, Wang, Zeng, & Liu, 2022) is used to track all the detected objects. We see that without image enhancement, we obtain an average FPS of 92.2. When adding the image enhancement model into the pipeline, the MBLLEN model is the slowest in terms of inference, whereas SCI is the fastest. However, we see that except for MBLLEN, all the models can support near-real-time inference on video input with the selected hardware.

Model	FPS
Low Light	92.2
EGAN-JT (mine)	25.7
ZDCE	40.5
IAT	23.6
MBLLEN	**5.6**
SCI	**71.3**

Table 4.14

Experiment 5 results. Comparison of inference speed using different enhancement models.

Figures 4.15 and 4.16 show the results of applying the proposed processing pipeline on video input using a subset of the image enhancement models.

In Figure 4.15, the image is taken from a nighttime driving video.[1] From the results displayed, we see that the illumination map is much smoother for the image processed by the EGAN-JT compared to the others. This can be easily observed around the street lights in the image. Also, processing the image using Zero-DCE and IAT results in some over-exposure. In the IAT image, the over-exposure can be seen around the road, whereas for the Zero-DCE image, the over-exposure is more globally distributed. Another thing to note is the dark regions in the image. The night sky in the image is relatively untouched

[1]Original video taken from https://www.youtube.com/watch?v=nABR88G_2cE

for all images except that processed by Zero-DCE, where we can clearly observe unevenness in the illumination map. The illumination levels are comparatively lower for the image processed by EGAN-JT, but the detection performance still seems to be the best for this model. The "low light" model pipeline, which represents the un-enhanced image, misses two objects in the image due to low light conditions. Images from the other models suffer from over-exposure and reduced contrast, and due to this, they also fail to detect all of the objects in the image.

Figure 4.15

Comparison of detection results in nighttime driving video.

In contrast to Figure 4.15, the images in Figure 4.16 are taken from a driving video captured in broad daylight.[2] It is important to observe the performance of the image processing pipeline under both conditions, so that we may deploy the models most effectively. Over-exposure is observed for the images processed by the Zero-DCE and IAT model pipelines. The illumination map is quite smooth for all the images, which is to be expected, as there are no dark regions in the original image. However we can clearly observe a shift in hue for the image processed using the Zero-DCE based model. In terms of detection performance, we can see that the "low light" model does not perform as well, since it has been trained on the un-enhanced ExDark dataset and is unable to generalize well to brighter images. This seems to be in alignment with the results presented by Loh and Chan (2019). The IAT model also fails to detect an object in the image due to over-exposure, which reduces contrast and blurs object boundaries.

[2]Original video taken from https://www.youtube.com/watch?v=IpEpTWIDL4Q

Figure 4.16

Comparison of detection results in daytime driving video.

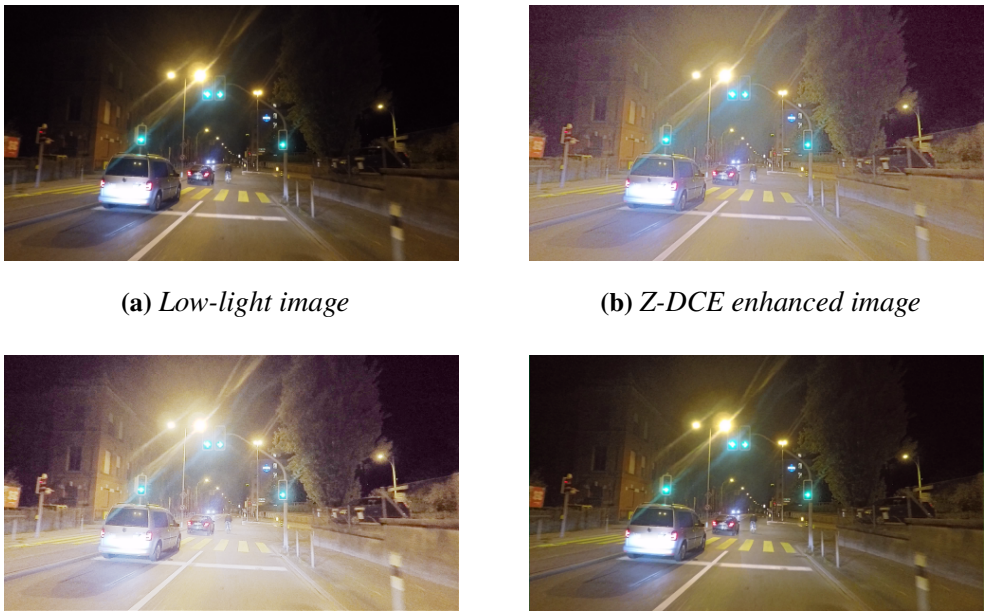
4.9 Image segmentation results on the ACDC dataset

The Adverse conditions dataset with correspondences (ACDC) (Sakaridis et al., 2021) is a large diverse dataset of images based on four conditions: nighttime, snow, fog, and rain. The dataset consists of 4006 images of size 1920×1080 divided equally among the four conditions. For the purpose of low-light perception, this study is concerned with the data pertaining to the nighttime images. There are 1006 nighttime images with 400 training images, 106 validation images, and 400 un-annotated test images. Since the ground-truth segmentation masks for the test set is not provided in the dataset, the numbers presented pertain to the validation set.

I fine-tuned the Deeplabv3-plus model (L.-C. Chen et al., 2018) on the 400 images in the ACDC training set. The backbone used for the Deeplabv3-plus model in Experiment 6 (Table 4.15) is Resnet-50. The model is initialized using weights pre-trained on the Cityscapes dataset (Cordts et al., 2016). It is then fine-tuned on the ACDC train set for 30 epochs. The MMSegmentation toolbox (Contributors, 2020) was used to carry out this experiment.

From the results in Table 4.15, it can be observed that the SCI (Ma et al., 2022) is the only model that improves upon the segmentation results when compared to the baseline of model trained on original images, and even then, it is a marginal improvement. EnlightenGAN-JT performs the worse out of all the models. Table 4.16 presents a break-

down of mIoU by class categories.



(a) *Low-light image*

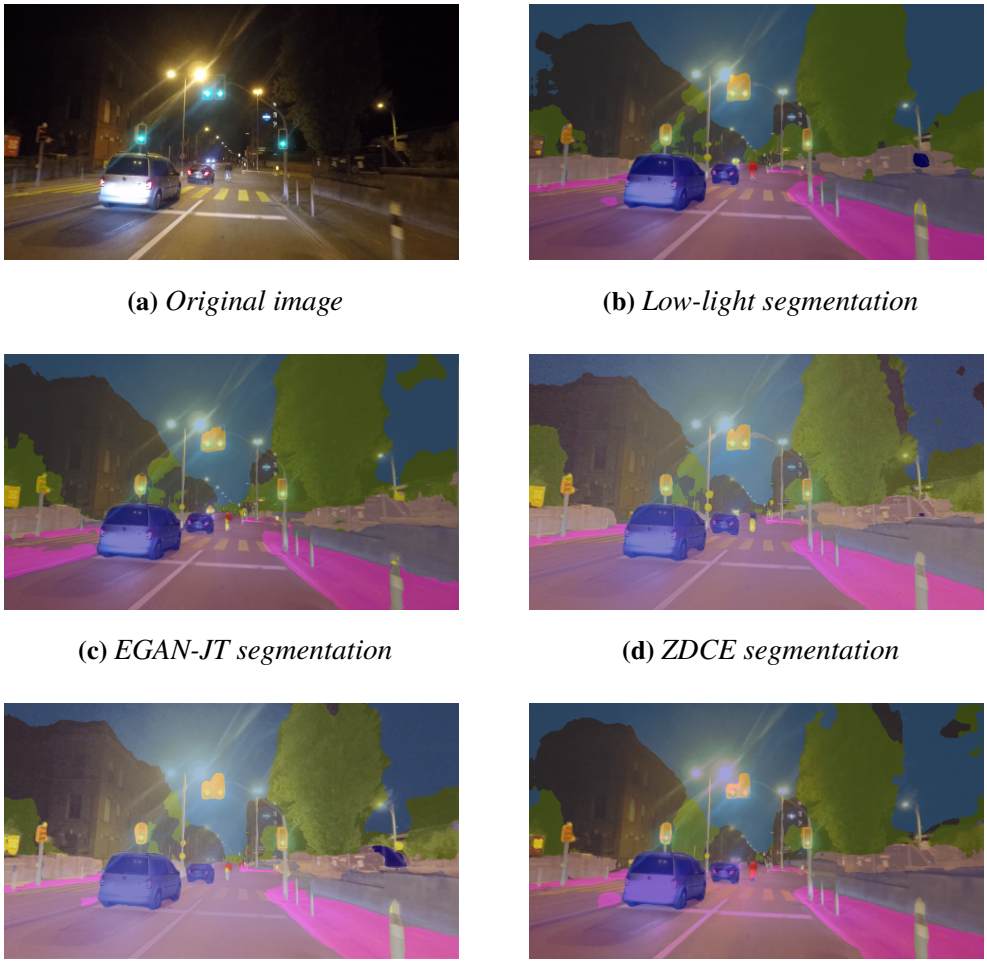
(b) *Z-DCE enhanced image*

(c) *SCI enhanced image*

(d) *EGAN-JT enhanced image*

Figure 4.17

Examples of image enhancement from ACDC dataset using various low-light enhancement models.



(a) *Original image*

(b) *Low-light segmentation*

(c) *EGAN-JT segmentation*

(d) *ZDCE segmentation*

(e) *SCI segmentation*

(f) *IAT segmentation*

Figure 4.18

Examples of Deeplabv3-plus semantic segmentation from the ACDC dataset using various low-light enhancement models.

Sr.	Model	Mean Accuracy	mIoU
1	Low Light	61.62	47.1
2	EGAN-JT (mine)	53.66	40.47
3	ZDCE	58.12	44.62
4	SCI	**62.60**	**47.93**
5	IAT	55.02	41.42

Table 4.15

Experiment 6 results. Evaluation of trained Deeplabv3-plus models on the ACDC dataset for the segmentation task. A breakdown by class is presented in Table 4.16.

4.10 Image segmentation results on the NightCity dataset

The NightCity dataset (Tan et al., 2020) is a dataset composed of 4297 images with pixel-level annotations for nighttime driving scene parsing. The ACDC dataset, although it contains images for nighttime driving scenarios, is not as diverse. The NightCity dataset is the largest dataset available for nighttime driving scene segmentation. It contains 2998 images for training and 1299 images for validation. The images are of size 1024×512. Figure 4.19 presents an example of an image and its enhanced versions from the NightCity dataset.

Deeplabv3-plus with the ResNet-50 backbone is fine-tuned on the NightCity training set and evaluated on the validation set. The model is initialized using weights pre-trained on the Cityscapes dataset (Cordts et al., 2016). It is then fine-tuned on the NightCity training set for 50 epochs. As in the previous experiment, the MMSegmentation toolbox (Contributors, 2020) was used in fine-tuning and evaluation.

Table 4.17 shows results for the scene parsing task after image enhancement using select low-light enhancement models. From the results, we can see that, unlike in the previous experiment with the ACDC dataset, the low-light image enhancement step results in a significant increase in accuracy on the image segmentation task. All the low-light enhancement models evaluated positively affect the mean IoU on the segmentation task, with increases as much as 3.22%.

EGAN-JT performs the best, which is interesting, as it was the worst performing model on the ACDC dataset. We can infer that the most appropriate low-light enhancement

Sr.	Class	Low Light	Z-DCE	SCI	EGAN-JT	IAT
1	Road	93.96	93.61	**94.07**	93.18	93.04
2	Sidewalk	**73.68**	70.90	72.63	70.31	67.15
3	Building	79.25	**79.47**	79.44	76.89	76.03
4	Wall	49.92	47.57	**52.13**	38.63	46.54
5	Fence	41.37	43.89	**46.71**	40.63	34.68
6	Pole	49.58	**49.77**	48.85	41.82	46.56
7	Traffic Light	53.20	49.85	**56.08**	40.46	52.24
8	Traffic Sign	40.89	36.71	**41.57**	30.57	30.33
9	Vegetation	70.19	70.33	**70.84**	67.95	67.80
10	Terrain	8.48	9.27	**12.82**	9.61	8.15
11	Sky	82.32	82.76	**83.68**	81.77	81.35
12	Person	**39.84**	27.64	39.04	30.43	33.69
13	Rider	12.78	11.85	**17.06**	3.87	8.97
14	Car	**64.79**	62.35	60.51	52.52	56.26
15	Truck	-	-	-	-	-
16	Bus	-	-	-	-	-
17	Train	70.21	69.27	**79.22**	68.30	57.68
18	Motorcycle	**17.93**	4.23	13.06	6.93	0.32
19	Bicycle	**46.57**	38.37	42.91	15.03	26.26

Table 4.16

Experiment 6 results by class.

model will not be the same across datasets and will in fact depend on the composition (size and diversity) of the dataset in question. Figure 4.18 shows the IoU results by class.

(a) *Low-light image*

(b) *Z-DCE enhanced image*

(c) *SCI enhanced image*

(d) *EGAN-JT enhanced image*

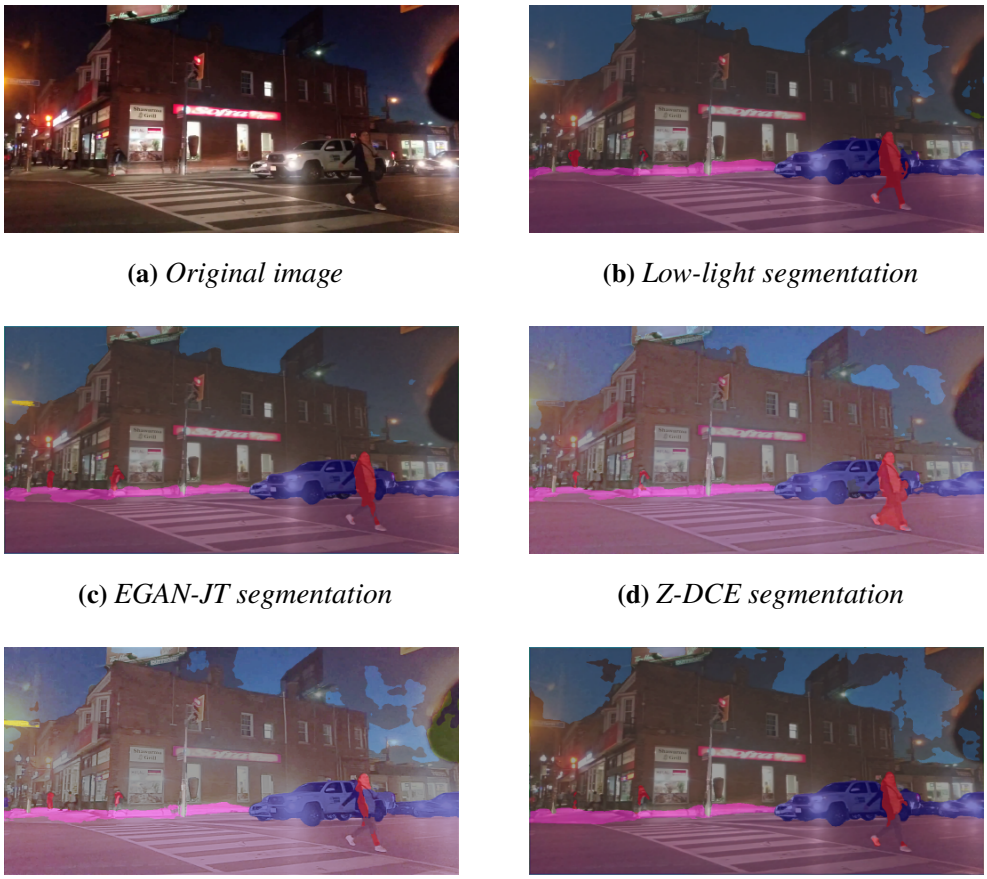
Figure 4.19

Examples of image enhancement from the NightCity dataset using various low-light enhancement models.

Sr.	Model	Mean Accuracy	mIoU
1	Low Light	59.11	46.78
2	EGAN-JT (mine)	**61.72**	**50.00**
3	ZDCE	60.11	48.17
4	SCI	59.18	48.52
5	IAT	59.72	48.92

Table 4.17

Experiment 7 results. Evaluation of trained Deeplabv3-plus models on the NightCity dataset for the segmentation task. A breakdown by class is presented in Table 4.18.



(a) *Original image*

(b) *Low-light segmentation*

(c) *EGAN-JT segmentation*

(d) *Z-DCE segmentation*

(e) *SCI segmentation*

(f) *IAT segmentation*

Figure 4.20

Examples of Deeplabv3-plus semantic segmentation from the Nightcity dataset using various low-light enhancement models.

Sr.	Class	Low Light	Z-DCE	SCI	EGAN-JT	IAT
1	Road	90.09	90.15	90.29	**90.46**	90.34
2	Sidewalk	49.93	49.68	50.65	**51.07**	50.86
3	Building	83.06	83.10	83.30	**83.64**	83.41
4	Wall	55.03	55.70	**57.16**	56.53	55.42
5	Fence	50.53	49.37	50.37	**51.89**	51.50
6	Pole	35.74	36.10	36.7	**36.92**	36.83
7	Traffic Light	**27.69**	25.82	26.24	27.36	25.42
8	Traffic Sign	48.85	52.41	51.23	**52.75**	52.29
9	Vegetation	58.28	58.41	58.93	**59.35**	59.35
10	Terrain	**26.26**	24.37	23.02	25.61	23.71
11	Sky	87.34	86.99	87.34	87.34	**87.46**
12	Person	48.03	46.73	48.91	**49.96**	49.52
13	Rider	1.78	1.61	0.37	**4.60**	1.16
14	Car	80.03	79.86	79.58	79.97	**80.54**
15	Truck	**60.71**	58.26	56.56	59.43	56.91
16	Bus	55.04	62.84	62.69	**66.66**	64.49
17	Train	3.84	26.65	30.22	**35.84**	30.94
18	Motorcycle	-	-	-	-	-
19	Bicycle	26.60	27.1	28.30	**30.70**	29.32

Table 4.18

Experiment 7 results by class.

CHAPTER 5
CONCLUSION

This study shows that the use of image enhancement models to pre-process images can improve the accuracy for the object detection task. We compare different low-light enhancement models to evaluate their efficacy. My experiments show that Zero-DCE and EnlightenGAN-JT are able to produce the best results, with a 1.4% improvement in mAP_{50} and a 4.6% improvement in $\text{mAP}_{50:95}$.

Further, some experiments were conducted to test the applicability of trained models over a range of input pixel intensity distributions. I simulate various lighting conditions by using gamma correction on the ExDark test dataset. I then used the trained models to evaluate the simulated test set. It can be concluded that EnlightenGAN-JT model is able to produce better results over a range of lighting conditions (gamma values). Also, by using model ensembling technique using all four (Low Light, EnlightenGAN, EnlightenGAN-JT and Zero-DCE) models, we can significantly improve the $\text{mAP}_{50:95}$ across the range of lighting conditions.

In further pursuance to this, I trained the YOLOv5s model on a dataset constructed by using multiple low-light enhancement models. The resulting model beats all the stand-alone low-light enhancement models in terms of efficacy for improving low-light object detection. It also suggests that multiple low-light image enhancement models enable learning of different sets of features to facilitate object detection.

We observe that the commonly used metrics for image enhancement models do not necessarily predict their utility for high vision tasks. We might assume that PSNR would be useful, as it provides a measure of the signal clarity in the output image. Similarly, SSIM provides a measure of the preservation of structural information in the output image. In terms of PSNR and SSIM, the Zero-DCE model is worse than MBLLEN, KinD, and EnlightenGAN. However, when utilized in a high-level vision task (object detection), it produces better results than any of the three models.

I trained the EnlightenGAN model jointly with the YOLOv5 model, which leads to an improvement over utilizing the pre-trained model. For this purpose, I omitted the self-feature-preservation loss (Jiang et al., 2021) from EnlightenGAN and instead backpro-

pogated the YOLO loss through the EnlightenGAN model during training. Using the resulting EnlightenGAN model produces images that are more color consistent with the original images, have comparatively lower illumination levels (than original EnlightenGAN), and have less noise, especially in the dark regions of the image. A similar experiment with self-feature-preservation loss is not able to produce competitive results.

Finally, I evaluated the utility of low-light enhancement models for improving performance on the semantic segmentation task. The experimental results (Section 4.10) show that for the NightCity dataset, all the low-light enhancement models help to improve the segmentation results, with EnlightenGAN-JT being the best. The evaluation results for the ACDC dataset (Section 4.9) are not as uniform. The SCI model seems to marginally improve the segmentation results, but other than that, other image enhancement models do not seem to help for this dataset. This may be because the ACDC dataset has fewer images and is not as diverse as the NightCity dataset.

The experiments conducted in this study lead us to conclude that use of low-light image enhancement models as a pre-processing step in training and inference for driving perception tasks, such as object detection and semantic segmentation, can provide significant improvements. In addition, for object detection task, joint training of low-light enhancement model with the YOLO loss provides noteable improvements. The vision pipeline proposed for object detection and tracking is able to run in near realtime, as shown through experiments, and thus can be operationalized with ease.

REFERENCES

Al Sobbahi, R., & Tekli, J. (2022). Comparing deep learning models for low-light natural scene image enhancement and their impact on object detection and classification: Overview, empirical evaluation, and challenges. *Signal Processing: Image Communication*, *109*, 116848. Retrieved from https://www.sciencedirect.com/science/article/pii/S092359652200131X doi: https://doi.org/10.1016/j.image.2022.116848

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2019). Yolact: Real-time instance segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9156-9165.

Bychkovsky, V., Paris, S., Chan, E., & Durand, F. (2011). 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.*

Chang, J.-R., & Chen, Y.-S. (2018). *Pyramid stereo matching network.* arXiv. Retrieved from https://arxiv.org/abs/1803.08669 doi: 10.48550/ARXIV.1803.08669

Chen, C., Chen, Q., Xu, J., & Koltun, V. (2018). Learning to see in the dark. In *2018 ieee/cvf conference on computer vision and pattern recognition* (p. 3291-3300). doi: 10.1109/CVPR.2018.00347

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European conference on computer vision.*

Chen Wei, W. Y. J. L., Wenjing Wang. (2018). Deep retinex decomposition for low-light enhancement. In *British machine vision conference.*

Contributors, M. (2020). *MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark.* https://github.com/open-mmlab/mmsegmentation.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213-3223.

Cui, Z., Li, K., Gu, L., Su, S., Gao, P., Jiang, Z., ... Harada, T. (2022). *You only need 90k parameters to adapt light: A light weight transformer for image enhancement and exposure correction.* arXiv. Retrieved from https://arxiv.org/abs/2205

.14871 doi: 10.48550/ARXIV.2205.14871

Cui, Z., Qi, G.-J., Gu, L., You, S., Zhang, Z., & Harada, T. (2021, October). Multitask aet with orthogonal tangent regularity for dark object detection. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)* (p. 2553-2562).

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

Forbes. (2021, July). Can tesla really do without radar for full self-driving? In *Forbes.* Retrieved from https://www.forbes.com/sites/jamesmorris/2021/07/03/can-tesla-really-do-without-radar-for-full-self-driving/?sh=38b2e9d77007

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*.

Guo, C. G., Li, C., Guo, J., Loy, C. C., Hou, J., Kwong, S., & Cong, R. (2020, June). Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (p. 1780-1789).

Guo, H., Lu, T., & Wu, Y. (2021). Dynamic low-light image enhancement for object detection via end-to-end training. In *2020 25th international conference on pattern recognition (icpr)* (p. 5611-5618). doi: 10.1109/ICPR48806.2021.9412802

He, K., Gkioxari, G., Dollár, P., & Girshick, R. B. (2017). Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980-2988.

Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., ... Wang, Z. (2021). Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, *30*, 2340–2349.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., ... Jain, M. (2022, November). *ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.7347926 doi: 10.5281/zenodo.7347926

Li, C., Guo, C., Han, L., Jiang, J., Cheng, M.-M., Gu, J., & Loy, C. C. (2021). Low-light image and video enhancement using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, C., Guo, J., Porikli, F., & Pang, Y. (2018). Lightennet: A convolutional neural net-

work for weakly illuminated image enhancement. *Pattern Recognition Letters*, *104*, 15-22. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167865518300163 doi: https://doi.org/10.1016/j.patrec.2018.01.010

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision.*

Liu, M.-Y., Breuel, T., & Kautz, J. (2017). *Unsupervised image-to-image translation networks.* arXiv. Retrieved from https://arxiv.org/abs/1703.00848 doi: 10.48550/ARXIV.1703.00848

Loh, Y. P., & Chan, C. S. (2019). Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, *178*, 30-42. doi: https://doi.org/10.1016/j.cviu.2018.10.010

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

Lv, F., Lu, F., Wu, J., & Lim, C. S. (2018). Mbllen: Low-light image/video enhancement using cnns. In *Bmvc.*

Ma, L., Ma, T., Liu, R., Fan, X., & Luo, Z. (2022). Toward fast, flexible, and robust low-light image enhancement. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5627-5636.

Rashed, H., Ramzy, M., Vaquero, V., Sallab, A. E., Sistu, G., & Yogamani, S. (2019). *Fusemodnet: Real-time camera and lidar based moving object detection for robust low-light autonomous driving.* arXiv. Retrieved from https://arxiv.org/abs/1910.05395 doi: 10.48550/ARXIV.1910.05395

Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*, 1137-1149.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, *abs/1505.04597*.

Sakaridis, C., Dai, D., & Van Gool, L. (2021, October). ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the ieee/cvf international conference on computer vision (iccv).*

Sasagawa, Y., & Nagahara, H. (2020). Yolo in the dark - domain adaptation method for

merging multiple models. In *Eccv.*

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, *abs/1409.1556*.

Tan, X., Xu, K., Cao, Y., Zhang, Y., Ma, L., & Lau, R. W. H. (2020). Night-time scene parsing with a large real dataset. *IEEE Transactions on Image Processing*, *30*, 9085-9098.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., & Li, Y. (2022). Maxim: Multi-axis mlp for image processing. *CVPR*.

Vu, D., Ngo, B., & Phan, H. (2022). *Hybridnets: End-to-end perception network.*

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2022). *Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors.*

Wu, D., Liao, M., Zhang, W., & Wang, X. (2021). *Yolop: You only look once for panoptic driving perception.*

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Álvarez, J. M., & Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural information processing systems.*

Xingang Pan, P. L. X. W., Jianping Shi, & Tang, X. (2018, February). Spatial as deep: Spatial cnn for traffic scene understanding. In *Aaai conference on artificial intelligence (aaai).*

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., ... Darrell, T. (2020, June). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Ieee/cvf conference on computer vision and pattern recognition (cvpr).*

Yurtsever, E., Lambert, J., Carballo, A., & Takeda, K. (2020, 03). A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, *PP*, 1-1. doi: 10.1109/ACCESS.2020.2983149

Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2022). Robust multi-object tracking by marginal inference. In *European conference on computer vision.*

Zhang, Y., Zhang, J., & Guo, X. (2019). Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th acm international conference on multimedia* (pp. 1632–1640). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/3343031.3350926` doi: 10.1145/3343031.3350926