NANYANG TECHNOLOGICAL UNIVERSITY

# Task-Oriented Low Light Enhancement

Wang Jingchao G2304141E

School of Computer Science and Engineering

2024

# NANYANG TECHNOLOGICAL UNIVERSITY

MSAI Master Project MSAI/19/001

# Task-Oriented Low Light Enhancement

Submitted by:
Wang Jingchao G2304141E

under the supervision of
Chen Change Loy

School of Computer Science and Engineering

2021

**Statement of Originality**

I hereby certify that the work embodied in this report is the result of original research (and/or research survey), is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

[Signature of student in this space]

Wang Jingchao G2304141E

Date: [13/03/2024]

**Supervisor Declaration Statement**

I have reviewed the content and presentation style of this report and declare that it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research (and/or research survey) and the writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accordance with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

[Signature of supervisor in this space]

Chen Change Loy

Date: [13/03/2024]

**Authorship Attribution Statement**

(A) This report **does not** contain any material from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

---

# Contents

# Abstract

In practical applications, images captured under low-light conditions often suffer from insufficient quality due to environmental constraints during capture or limitations in camera performance, posing challenges for subsequent image analysis, recognition, and application. Low-light image enhancement (LLIE) is one of the key techniques to solve this issue, attracting widespread research. However, most low-light image enhancement methods focus on refining the visual fidelity of low-light images to meet the criteria of human visual perception. I believe that enhancing the performance of enhanced images in downstream tasks is the issue I must consider and can also be the way to refine the upstream model to fit specific downstream vision tasks. Therefore, I proposed a joint training framework that combines the information back-propagated from the downstream task with the existing loss function of the upstream model to refine the LLIE method. Specifically, the output enhanced image of the upstream is fed into downstream models, and loss values from the downstream model are returned, which are integrated with the upstream model to direct the update of the upstream model because I do not want to improve the downstream itself but to enhance the low-light enhancement approach. I use Zero-DCE++ [1] as the upstream model and Tinaface [2] as the downstream model to conduct experiments. To generate a low-light dataset, based on the WIDERFACE dataset [3], I utilize the low-light data generation pipeline EC-Zero-DCE [4] to prepare the training and test dataset. I demonstrate that enhanced images of my framework achieve better performance on downstream vision tasks than a single LLIE model.

**Keywords:** Low Light Enhancement, Joint Training, Backpropagation.

# Acknowledgement

With the help of teachers and PhD students over the two semesters, I was able to complete this project. First and foremost, I would like to express my appreciation to my supervisor, Professor Chen Change Loy, who provided me with the opportunity for this project. From understanding the topic selection, subsequent research, to the implementation of code and experimental evaluation, none of these could have been accomplished without the careful attention, patient guidance, and timely supervision of Professor Chen Change Loy. Since I embarked on this project, Professor Chen Change Loy has offered me invaluable guidance with his comprehensive and meticulous teaching attitude, as well as his rigorous research spirit. I would like to express my sincere gratitude to him once again.

Secondly, I would like to thank Mr. Zhang Junzhe for all the help he has provided me. Mr. Zhang Junzhe assisted me in understanding relevant knowledge and continuously mentored me in completing the project. Not only did Mr. Zhang Junzhe provide me with guidance in learning and practical opportunities, but he also offered valuable advice for my future direction. Under his guidance, I have not only acquired professional knowledge but also accumulated experience in academic research and project management. Thank you for your assistance.

I also want to express my gratitude to all the classmates and teachers who have helped and taught me throughout this project. Thank you all for everything you have done. Finally, I would like to extend my sincere thanks and best wishes to Nanyang Technological University for nurturing me and providing me with learning opportunities.

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Background

In real-world applications, due to limitations in the environment or devices, images captured under low-light conditions often suffer from issues such as insufficient lighting, low contrast, and loss of details. These challenges can significantly hinder subsequent image processing tasks. Low-light image enhancement (LLIE) emerges as a crucial technology to tackle this problem. It aims to improve the quality and visualization of low-light images, thereby enhancing the accuracy and reliability of subsequent image processing and analysis tasks.

Low-light image enhancement techniques is applied in various fields, including surveillance systems, medical imaging, autonomous driving, night photography, etc. Traditional enhancement methods may include Histogram Equalization [5], filtering [6], contrast enhancement [7] techniques, and Retinex model-based methods [8] [9] [10] [11] [12] [13].

In recent years, with the development of deep learning technology, neural network-based methods have become the mainstream in the field of low-light image enhancement. In most cases, these methods achieve better accuracy and robustness. Most of these works, as supervised learning models, aim to optimize the appearance distance between the output and the ground truth. Although there are also unsupervised learning models, like EnlightenGAN [14] and Zero-DCE [15], they also focus on improving the effect of low-light images themselves.

However, in most cases obtaining the images with enhanced illumination is not the final purpose. Instead, I apply low light image enhancement as preparation work for subsequent image processing tasks. Utilizing the feedback from downstream tasks to improve the upstream LLIE model is the mission in my work.

## 1.2 Motivation

As discussed in the previous section, most LLIE works overlook the performance of downstream image processing tasks. The loss from the training process and the evaluation experiments only focus on enhanced images. Some works [14] have acknowledged the necessity to incorporate downstream tasks, but they merely do so as an evaluation experiment. In my

work, I investigate a framework to apply joint training of low-light enhancement methods along with the downstream models. This framework utilizes information propagated backward from downstream models to enhance the low-light enhancement approach.

## 1.3 Objectives and Specifications

In this project, I focus on both the upstream model, which consists of low-light enhancement methods, and the downstream model, which involves subsequent computer vision tasks utilizing enhanced images. However, I do not pay attention to refining the downstream model itself. Instead, my efforts are devoted to enhancing the low-light enhancement approach. To achieve this, I introduce the downstream loss to the upstream model to refine it alongside the initial upstream loss.

(a) Low Light Image

(b) Zero-DCE

(c) Joint Training

(d) Original Image

Figure 1.1: A comparison on the visualizing human face recognition results of different models. (a) Low light images. (b) The image enhanced by only Zero-DCE++ model and then recognized by Tinaface. (c) The images ehanced by my joint training models. (d) Original normal light images

In my experiments, I utilized Zero-DCE++ [1] as the upstream model and Tinaface [2] as the downstream model to conduct face recognition tests on low-light images. As shown in Fig. 1.1, low-light images (Fig. 1.1(a)) pose significant challenges for face recognition tasks, resulting in partial recognition results as shown in Fig. 1.1(d). Compared with the enhanced image generated by Zero-DCE++ (Fig. 1.1(b)), the result produced by my model (Fig. 1.1(c)) not only achieves higher confidence in recognition but also successfully identifies parts that Fig. 1.1(b) failed to recognize.

## 1.4 Major contribution of the Dissertation

In summary, my work's contribution can be summarized in 3 aspects.

1. I proposed a task-oriented joint training framework to improve the LLIE model with the information backpropagated from the subsequent computer vision model.
2. I generated low light images based on WIDERFACE [3] dataset with the data generation pipeline.
3. I demonstrated that my method get better evaluation result on the downstream vision model than LLIE model only.

## 1.5 Organisation of the Report

The organizational structure of this article is as follows. In chapter 2, I briefly introduce the related work of my model, which consist of traditional LLIE methods and LLIE methods considering downstream vision tasks. In chapter 3, I specifically introduced how my framework refines the LLIE model based on both upstream and the loss information backpropagated from downstream models. The evaluation results of the my framework are presented in Chapter 4. Chapter 5 summarizes the entire text and provides recommendations for further work.

# Chapter 2

# Literature Review

## 2.1 Low Light Images Enhancement

### 2.1.1 Low-Light Images Enhancement for Images

Currently, deep learning-based low-light image enhancement methods have become the mainstream in the field of LLIE. Most deep learning-based LLIE methods focus on improving the visual quality of images themselves, prioritizing features that satisfy human visual perception. [16] [8].

#### Supervised Methods

Among these deep learning methods, supervised learning-based LLIE methods comprise the vast majority. LLNet [17] was the first LLIE approach to employ deep learning techniques. It proposed a method for enhancing natural low-light images using a class of deep neural networks called Stacked Sparse Denoising Autoencoder (SSDA). In addition to LLNet trained with darkened images and noisy images, [17] also proposed staged LLNet (S-LLNet), which consists of independent modules for contrast enhancement in stage 1 and denoising in stage 2. Furthermore, [17] also proposed a method for generating training data by synthetically modifying images in the network database to simulate low-light conditions, which inspired the development of low-light image generators in the future.

Another significant CNN-based LLIE method is MBLLEN [18], which transforms the low-light enhancement task into an end-to-end training process. MBLLEN consists of three modules: the feature extraction module (FEM), enhancement module (EM), and fusion module (FM). FEM extracts image features from different convolutional layers. It has a unidirectional 10-layer network structure with 32 convolutional kernels of size 3x3 and a convolution stride of 1, followed by ReLU activation functions. EM consists of the same number of sub-networks as FEM's convolutional layers. Each sub-network applies enhancement processing separately to enhance the image features from multiple aspects. Finally, FM merges all the images output from the EM sub-networks and obtains the final enhanced image by convolving with 3-channel 1×1 convolutional kernels. MBLLEN also proposes a new loss function, which consists of structure loss, context loss, and region loss.

Lv et al. [19] propose another end-to-end attention-guided method based on a multi-branch convolutional neural network, which learns two attention maps to guide low-light enhancement and denoising, respectively. The fully convolutional network can be divided into four sub-networks: an Attention Net, a Noise Net, an Enhancement Net, and a ReinforceNet. The Attention Net is a UNet designed to guide the Enhancement Net to correctly enhance the under-exposed regions of the input image and prevent excessive enhancement in normally exposed regions. To distinguish between noise and image textures and avoid introducing blurring in the enhanced image, the Noise Net estimates the noise distribution before adapting the denoising process. Based on the assumption that the noise distribution is closely linked to the exposure distribution, the Noise Net utilizes the UE-attention map from the Attention Net to assist in generating a noise map. The key point of the Enhancement Net is to break down the low-light enhancement task into multiple sub-missions. Like MBLLEN [18], the Enhancement Net also consists of FEM, EM, and FM. Finally, the ReinforceNet is designed to overcome the low-contrast drawback and improve the details.

LEDNet [4] is designed to address two problems: low-light enhancement and deblurring, using an encoder-decoder structure. The low-light enhancement encoder consists of three blocks, each composed of a Residual Block, a Residual Downsampling Block [20], a Pyramid Pooling Module (PPM) [21], and a Curve Non-Linear Unit (CurveNLU). LEDNet introduces PPM [21] to remove local noise from the enhanced image by incorporating global contextual priors into the network. Inspired by Zero-DCE [15], the Curve Non-Linear Unit is proposed to conduct feature transformation by learning a non-linear activation function. The deblurring decoder consists of three convolutional blocks, each containing two Residual Blocks, one Residual Upsampling Block [20], and a FAC Layer [22] to link the enhanced features from the encoder to the decoder. Inspired by FAC [22], LEDNet proposed the Filter Adaptive Skip Connection (FASC) to utilize the enhanced information from the encoder to aid in deblurring.

Apart from directly training an end-to-end network, utilizing the Retinex theory [8] is another direction for low-light enhancement. The Retinex theory suggests that an image I is composed of an illumination image and a reflectance image. The illumination component describes the external lighting environment, denoted as L, while the reflectance component refers to the intrinsic characteristics of objects themselves, denoted as R. Based on this theory, the following mathematical model can be constructed:

$$I = R * L \tag{2.1}$$

The key point of low-light image enhancement based on the Retinex theory is the rational decomposition of the image S into R and I and correcting the low light illumination component I to normal light llumination component $\hat{L}$. Then we can obtain enhanced image $\hat{I}$ through the equation:

$$\hat{I} = R * \hat{L} \tag{2.2}$$

To achieve this, RetinexNet [9] proposed a decomposition module called DecomNet and an illumination adjustment module called EnhanceNet. DecomNet decomposes the low-light image into illumination and reflectance components and EnhanceNet complete the light enhancement by transforming illumination component L into $\hat{L}$ and denoising the reflectance component R to $\hat{R}$. Finally, the enhanced image can be calculated with $\hat{I} = \hat{R} * \hat{L}$.

Following Retinex-Net [9], KinD [23] believes that low-light images also suffer from image degradation in addition to low illumination, but the degradation is covered by the low illumina-

tion component. Therefore, KinD [23] refine the equation(2.1) as followed:

$$I = R * L + E = R * (L + \tilde{E}) \tag{2.3}$$

Where E is the degradation component and $\tilde{E}$ is the degradation of the decomposed reflectance component. In this case, after decomposition into R and L, kinD [23] not only correct the illumination component L into $\hat{L}$, but also correct reflectance component R to get ideal reflectance component $\tilde{R} = R - \tilde{E}$. And the final result is $\hat{I} = \tilde{R} * \hat{L}$. KKinD [23] Network can be decomposed into three sub-networks: Layer Decomposition Net, Reflectance Restoration Net, and Illumination Adjustment Net. The Layer Decomposition Net has two branches, one for predicting the reflectance component and the other for predicting the illumination component. The llumination Adjustment Net correct the illumination component into $\hat{L}$. The Reflectance Restoration Net estimate the corrected reflectance component $\tilde{R}$. The authors suggested that noise is a complex distribution within the reflectance image R, and its distribution largely depends on the distribution of illumination L. So the Reflectance Restoration Net uses both R and L as input to restore the reflectance component. It is also worth noting that the Reflectance Restoration Net uses the reflectance component R from the normal light image as ground truth since the reflectance component of low-light images contains more degradation components. KinD++ [24] improved the Reflectance Restoration Net in advance by introducing multi-scale illumination attention module(MSIA).

[25] propose a more complex network which introduces an encoder-decoder network and uses a spatially variant recurrent neural network (RNN) as an edge enhancement stream. LR3M [25] enhances the low-light image and denoises jointly in a sequential manner. Since [25] believes that simultaneously estimating the illumination and reflectance maps alternately could result in residual noise in both the illumination and reflectance maps, LR3M [25] sequentially estimates both the illumination and reflectance maps. After the estimation of the illumination map independent from the reflectance map, while improving the reflectance map based on both the enhanced illumination and the original low light image, this network also utilizes low-rank prior to constrain the reflectance map to remove the noise.

Recently, Wu et al. [26] proposed URetinexNet, which is a deep unfolding network based on the Retinex theory. URetinexNet [26] unfolds the optimization problem of solving R and L into a learnable network. By formulating the decomposition problem as an implicit prior regularization model, URetinexNet [26] introduces three learnable modules: Initialization module, Unfolding optimization module, and Illumination adjustment module. Initialization module returns the initialized illumination and reflectance components while preserving the important information of the input image. Unfolding optimization module unfolds an optimization problem into a learnable network, adaptively fitting the implicit prior in a data-driven manner, and finally achieves noise suppression and detail preservation. And the Illumination adjustment module flexibly enhances the illumination through user-defined ratios.

#### Unsupervised Methods

Supervised methods require paired training images, which result in limited generalization and perform poorly in real-world scenarios [27]. To address this issue, unsupervised LLIE methods are introduced. GAN-based methods are a popular approach to establish non-paired mappings between low-light and normal-light image spaces without relying on precise matching of images. GAN consists of two models: the generative model and the discriminative model. The

task of the generative model is to generate enhanced images that appear natural and similar to the original low-light image. The task of the discriminative model is to determine whether a given instance looks natural or artificially generated.

EnlightenGAN [14] introduces an attention-based U-Net [28] as the generator and a dual discriminator for global and local enhancement. To utilize paired training data, a method utilizing self-regularized perceptual loss is proposed to constrain the feature distance between low-light input images and enhanced images, which is adopted for both local and global loss by the GAN network.During the training process of GAN, a global discriminator often cannot handle spatially varying illumination in images, and several local regions require adaptive enhancement. To address this problem, EnlightenGAN proposes a global-local discriminator structure, which utilizes PatchGAN for discrimination.The local discriminator randomly crops local patches from both the output and real normal-light images and learns to distinguish whether they originate from real images or from the enhanced output. The generator standardizes the illumination channel I of the input RGB image to the range [0,1]. Then, it computes the self-regularization attention map with 1-I. Subsequently, the attention map is resized to match each feature map and applied by element-wise multiplication to all intermediate feature maps as well as the output image.

[29] propose an implicit Neural Representation method for Cooperative low-light image enhancement called NeRCo, which is a CycleGAN based method introducing the combination of text-driven appearance discriminator and CLIP. NeRCo introduces controllable fitting capability of neural representation to the low-light image enhancement task. Furthermore, NeRCo also introduces multi-modal learning to low-light image enhancement to learn diverse features from the efficient vision-language priors.

Zero-DCE [15] and Zero-DCE++ [1] proposed a new learning strategy called zero-Reference learning, which transforms light enhancement into an image-specific curve estimation task. To enable zero-reference learning, Zero-DCE [15] proposes a set of loss functions to maintain exposure or color information, making it easily extendable to generic lighting adjustments. Zhou et al. [4] utilize this characteristic to introduce a data synthesis pipeline EC-Zero-DCE to generate low-light images. We would also use this method to generate a darken dataset for experimentation.

### 2.1.2 Low-Light Images Enhancement for Downstream Vision Tasks

The LLIE approaches above emphasize enhancing the quality of the images themselves. Several works also focus on enhancing images to improve downstream vision tasks. Simply applying the LLIE method as a preprocessing step could lead to unsatisfactory results [30]. Recently, another direction has been proposed that explores end-to-end pipelines, optimizing both enhancement and individual tasks during training, and that is what my work does.

Liang et al. [31] proposed a Recurrent Exposure Generation (REG) module, which can be integrated with a Multi-Exposure Detection (MED) module to suppress non-uniform illumination and noise in low-light face detection tasks, aiming to improve face detection performance with low-light images. The REG module generates light-enhanced images while encoding historical regional information to simulate non-linear multi-exposure processes within the camera. These pseudo-exposures are then inputted into the MED module to generate face bounding boxes. REG and MED are linked together to form a whole end-to-end framework.

To conduct face detection without low-light annotations, Wang et al. [32] proposed a joint High-Low Adaptation (HLA) framework utilizing Bidirectional low-level adaptation and multitask high-level adaptation for low-light face detection. For the low-level image enhancement part, HLA degrades the quality of normal-light images, thus bringing both domains closer together to improve the enhanced image. For high-level tasks, HLA employs a combination of context-based and contrastive learning techniques to effectively align features across different domains.

With the emergence of real-world low-light images [33–35], some works focus attention on object detection under low-light conditions. Zhang et al. [36] proposed the IA-YOLO approach to determine the best setup for the filters used in the differential image processing module. This framework makes images adaptively enhanced to improve detection performance. Specifically, Wang et al. proposed an image processing (DIP) module to deal with the challenging YOLO object detector in bad weather conditions and predict the parameters of the DIP module through a small convolutional neural network model (CNN-PP). The authors chose the one-stage detector YOLOv3 as the target detection network. The CNN-PP model and YOLOv3 learn appropriate DIP through end-to-end training, thus enhancing images for subsequent detection in a weakly supervised manner.

MAET [37] explores physical noise models of sensors and the image signal processing (ISP) pipeline to encode the intrinsic structure and then decode bounding box coordinates and categories for object detection. MAET adopts a self-supervised approach with a physical noise model and ISP to learn the intrinsic visual structure by encoding and decoding real-world illumination degradation transformations. MAET decodes the bounding box coordinates and classes for object detection. To avoid the over-entanglement of the two tasks, MAET imposes an orthogonal tangent rule to release the entanglement of objects and reduce features. This forms a parameter manifold, along which multi-task predictions can be geometrically represented by maximizing orthogonality between tangents along the outputs of each task.

Unfortunately, both [36] and [37] require large-scale data to achieve good results and did not decouple from upstream and downstream tasks. Recent work has started to consider high-level downstream computer vision tasks, like semantic segmentation [38]. [39] proposed a cascaded architecture to handle low-light enhancement and semantic-related tasks simultaneously. To enhance the collaboration between both upstream and downstream tasks and enable mutual guidance, they introduce a contrastive-alternative learning strategy to train the model parameters. This approach significantly enhances the representational capacity of the cascaded architecture.

However, the models mentioned above focus on dealing with specific downstream tasks. Inspired by vision-based backbone networks [40], Hashmi et al. [30] enhance hierarchical features through jointly optimizing feature enhancement and downstream tasks to improve several downstream vision tasks. Contrarily, I focus on improving LLIE approaches themselves. I utilize the detecting results backpropagated from the downstream model and combine them with the loss values of the upstream model to refine the LLIE approaches to improve the performance of the downstream tasks.

## 2.2 Object Detection

### 2.2.1 Generic Object Detection

Object Detection is a computer vision task aimed at accurately identifying objects in images or videos and determining their positions and categories. With AlexNet [41] winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012, deep learning methods have become the mainstream in the object detection field. R-CNN [42] proposed a two-stage detection framework, utilizing region proposals and CNN features for object localization. Fast R-CNN [43] and Faster R-CNN [44] enhanced both speed and precision by integrating feature extraction, proposal extraction, bounding box regression (region refinement), and classification into one single network. Mask R-CNN [45] extended Faster R-CNN [44] by incorporating an additional branch to concurrently predict object masks alongside the existing branch dedicated to bounding box detection. [46] introduces Feature Pyramid Network into object detection, enhancing the accuracy of object detection with Faster R-CNN [44]. SSD [47] introduced a one-stage object detection approach, predicting object categories and bounding box offsets directly from feature maps at multiple scales. YOLOs [48–50] are another series of one-stage approaches that transform object detection as regression problems forecasting bounding box coordinates and class probabilities in a single traversal of the network.

Returning to deep convolutional networks, after AlexNet [41], GoogLeNet [51] introduced the Inception structure to use different numbers of small filters, integrating feature information at different scales. ResNet [52] illustrates the significance of preserving the original information flow and introduces skip connections to address performance degradation in deeper networks. VGG [53] uses a sequence of consecutive 3x3 convolutional kernels instead of larger convolutional kernels.

With the advent of the Transformer [54], attention mechanisms have been introduced into the field of object detection. Relation Net [55] is an improvement over Faster R-CNN [44], which generates a series of regions of interest and then inputs these regions of interest into the Transformer to integrate information among different regions, thereby achieving feature enhancement. DETR [56], considering object detection as a set prediction problem, is the first work to utilize Transformers to accomplish object detection tasks.

### 2.2.2 Face Detection

As an application of generic object detection, the advancement of face recognition coincides with the progress of object detection. WIDERFACE [3] has been the most popular and challenging face detection benchmark since its proposal, leading to rapid advancements in face detection. Nearly all recent face detection methods derive from existing generic object detection methods. For example, based on SSD [47], S3FD [57] is a face detection SSD with scale invariance and a scale-adjusted anchor matching strategy. DSFD [58], a Dual Shot face detection network, introduces a Feature Enhancement Module and Improved Anchor Matching. Based on RetinaNet [59], RetinaFace [60] consists of five extra modules: Selective Two-step Regression, Selective Two-step Classification, Scale-aware Margin Loss, Feature Supervision Module, and Receptive Field Enhancement. Tinaface [2] is another simple but powerful RetinaNet-based network, which is the downstream model in my project.

# Chapter 3

# Approach

In this chapter, I introduce the implementation of the joint training low-light enhancement method and how I generate the low-light dataset. In the first section, I will introduce the generation pipeline EC-Zero-DCE [4]. In the second section, I will provide separate introductions for the upstream model, downstream model, and the guidance of the downstream model on the training of the upstream model.

## 3.1 Low Light Images Generated by EC-Zero-DCE

I design the framework to cater to different downstream vision tasks in low-light conditions, which require various types of datasets. Although some real datasets such as DARK FACE [34] and ExDark [33] have been proposed to satisfy specific low-light vision tasks, they are not suitable for flexible downstream tasks. In my project, I utilize synthetic low-light samples generated based on existing detection and segmentation datasets as training sources. I use the generation pipeline EC-Zero-DCE [4] to accomplish this task.

(a) Low Light (b) Normal Light (c) Over Exposure

Figure 3.1: Images with different exposure settings.

Zero-DCE models estimate light-enhancement adjustment curves to simulate normal light, allowing for pixel-wise adjustments for image enhancement. Consequently, reversed pixel-wise curves can also be utilized to simulate low light conditions. Additionally, as an unsupervised LLIE method, Zero-DCE [15] employs a set of differentiable non-reference losses to maintain spatial and color information. It adopts exposure control loss to measure the distance

between the exposure level of local regions and the preconfigured exposure level E, facilitating the resetting of the generating strategy. This is how EC-Zero-DCE [4] is reformulated into an Exposure-Conditioned variant.

EC-Zero-DCE [4] outputs low light images by controlling the low light exposure level. I can replace the well-exposedness value E with a low light value to darken the images. Following the approach of Zero-DCE, EC-Zero-DCE similarly adjusts the exposure level to generate images under different light conditions by setting different light values.

## 3.2 Joint Training Framework

In this section, I will introduce the task-oriented joint training framework. My objective is to improve both the visual perception quality and the effectiveness of subsequent vision tasks for low-light images. Within this framework, I address both the low-light enhancement model, regarded as the upstream model, and the subsequent computer vision model, which serves as the downstream model.

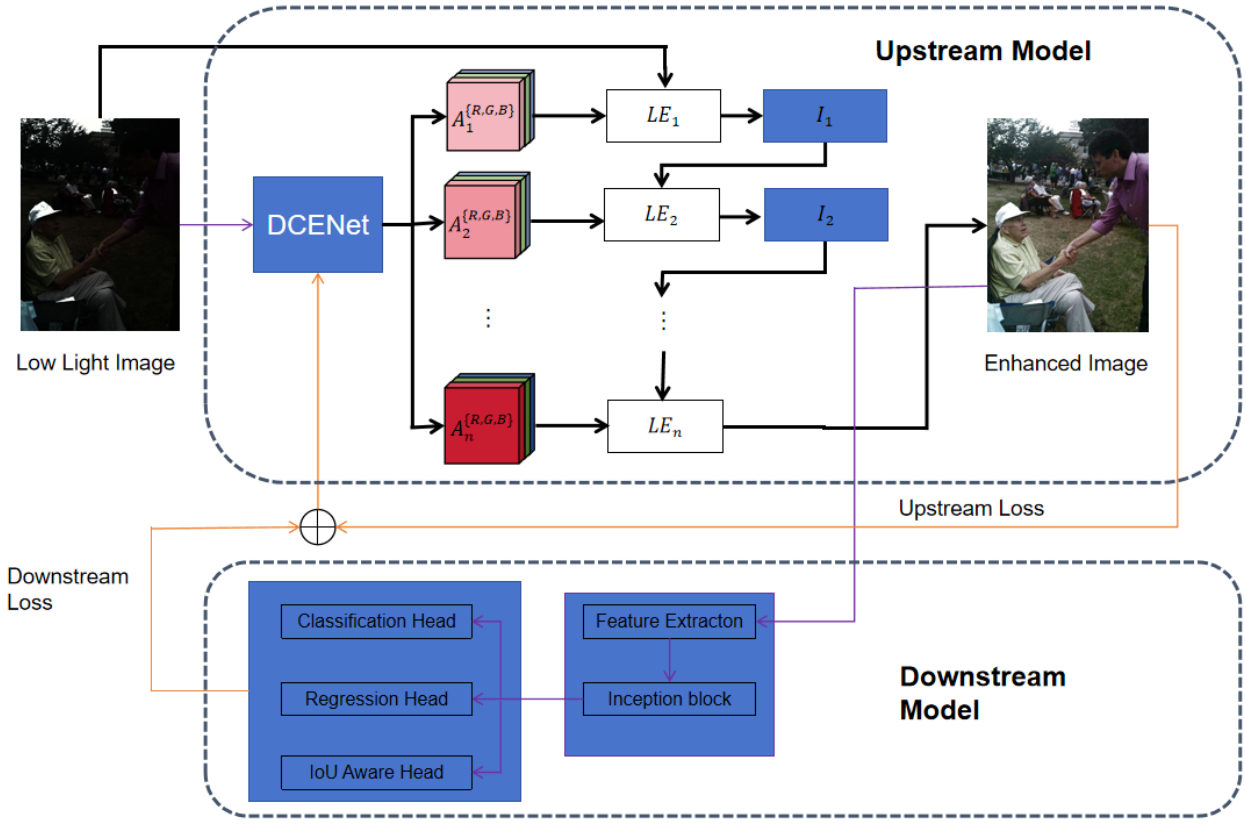



Figure 3.2: Overview of my approach.

### 3.2.1 The Overall Process

In this project, my focus lies on low-light face detection tasks. I utilize the Tinaface model [2] as the downstream model to perform face detection tasks using enhanced images and return

loss values. As for the upstream LLIE model, I opt for Zero-DCE++ [1] to generate enhanced images.

My objective is to enhance the performance of the LLIE upstream model by treating it as the central model during training. The output of the downstream loss functions serves a role similar to content loss [61] [62], combined with the upstream loss to refine the upstream model. The structure of my project is illustrated in Fig. 3.2

### 3.2.2 Downstream Model

Tinaface [2] is a powerful yet straightforward deep learning model crafted for face detection and recognition. Built upon the foundation of RetinaNet [59], Tinaface [2] discards specialized modules and techniques tailored specifically for face detection, which can be challenging to replicate and apply. It's noteworthy that Tinaface [2] treats face detection as a form of generic object detection, leveraging existing technologies in this domain. Its simplicity and scalability render Tinaface [2] an ideal downstream model for my project.

Compared with RetinaNet [59], Tinaface [2] incorporates several modifications, as outlined below:

1. Changed Batch normalization (BN) [63] to group normalization (GN) [64]. GN [64] is preferred over BN [63] due to its ability to mitigate the inaccuracies in batch statistical estimation, which can lead to a decline in model performance with decreasing batch sizes. GN [64] effectively removes the dependence on batch size, making it a more robust choice for normalization in various scenarios.
2. Introduced IoU-aware Branch [65], a single convolutional layer forecasting the Intersection over Union (IoU) between the identified box and the associated ground-truth object. This addition addresses the discrepancy between classification scores and localization accuracy in a single-stage object detector, improving overall performance.
3. Embedded DCNv1 [66] in the backbone architecture to further enhance the capacity of the Tinaface [2] model. Traditional convolution operations make strong assumptions about fixed and rigid sampling positions, limiting the network's ability to learn or encode intricate geometric transformations. DCNv1 [66] addresses this limitation by introducing deformable convolution operations, allowing the network to adaptively adjust the sampling positions based on learned offsets. This enhances the model's capability to capture spatial details and improve performance in object detection tasks.
4. Changed the bounding box regression loss from Smooth L1 Loss [43] to DIoU loss [67] to better align with IoU loss. Smooth L1 Loss [43] optimization target is not consistent with the regression evaluation metric IoU, as lower loss does not necessarily mean higher IoU. DIoU loss [67] addresses this issue by providing a more direct optimization target for IoU. Additionally, DIoU loss [67] enables the model to better handle small faces, which is a significant challenge in face detection tasks, particularly in datasets like WIDERFACE [3].

I designate Tinaface [2] model as $model_{down}$, which takes the output $I_{output}$ from the trained upstream model as input to obtain the downstream loss $L_{down}$. This downstream loss is transformed from the output loss value of the model to conform to the format and size of the loss

function of the upstream model. The downstream loss function is:

$$L_{down} = model_{down}(I_{output}) \tag{3.1}$$

### 3.2.3 Upstream Model

#### Light-Enhancement Curve

Inspired by the exposure adjustment curve commonly found in photo editing software, Zero-DCE [15] enhances low-light images by learning pixel-wise curves exclusively from the input low-light image, mapping it to a normal-light image. The pixel values of the images are normalized to the range [0, 1].

The curve that maps the low-light image to the enhanced image should adhere to several principles:

1. As the exposure values range from 0 to 1, to ensure an invariant range, the curve should include the points (0, 0) and (1, 1).
2. To maintain the relative exposure level of the image, the curve should be monotonically increasing within the interval [0, 1].
3. The curve equation must be simple to ensure differentiability.

Zero-DCE [15] employs the following formula to describe the curve:

$$LE(I(x);\alpha) = I(x) + \alpha I(x)(1 - I(x)) \tag{3.2}$$

In the formula, x is pixel coordinates and $\alpha \in [0,1]$ is the learnable parameter. $I(x)$ is the input image and $LE(I(x);\alpha)$ is the enhanced image.

This is a very ingenious curve formula. However, while $\alpha$ can adjust the curve, the curve is still essentially a quadratic function, which means the variation of the curve is not diverse enough. In order to fit more complex curves, a simple and effective approach is proposed: iteratively nesting this function. which means:

$$LE_1(x) = I(x) + \alpha_1 I(x)(1 - I(x)) \tag{3.3}$$

$$LE_n(x) = LE_{n-1}(x) + \alpha_n I(x)(1 - LE_{n-1}(x)) \tag{3.4}$$

By initiating iterative nesting with function (3.2), I can represent higher-order functions adhering to the principles. In each iteration, a new parameterr $\alpha_n$ is needed. I set the maximum value of n to be 8, indicating a total of 8 iterations.

However, it is unreasonable for the light enhancement level to be the same for every part of the images. In bright areas, I aim for a smaller degree of light enhancement, which means a smaller value for $\alpha$, possibly approaching 0. Conversely, in dark areas, I desire a greater degree of light enhancement, resulting in larger values for $\alpha$. In this case, $\alpha$ is not just a parameter but a function of pixel coordinates x, expressed as $A(x)$.

The final formula [15] can be expressed as:

$$LE_1(x) = I(x) + A_1(x)I(x)(1 - I(x)) \tag{3.5}$$

$$LE_n(x) = LE_{n-1}(x) + A_n(x)I(x)(1 - LE_{n-1}(x)) \tag{3.6}$$

Where $A_n(x)$, whose range is [-1,1], is the learnable parameters which define the pixel-wise light enhancement functions.

### DCENet

To learn the best curve parameters, the Deep Curve Estimation Network (DCE-Net) is proposed. The DCENet takes a low-light image as input and produces a collection of pixel-wise curve parameter maps representing higher-order curves corresponding to the input. Since the data to be fitted is not very complex, the network is very simple, consisting of 7 layers: 6 hidden layers and 1 output layer.

The structure of DCENet are shown in Fig 3.3.

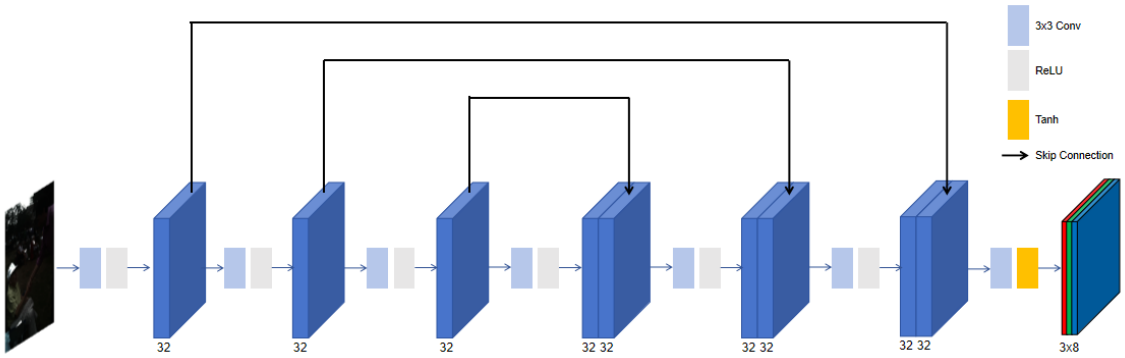



Figure 3.3: The architecture of DCENet.

The convolutional layers consist of 32 convolutional kernels, each with a size of 3x3 and a stride of 1. To maintain the correlation between neighboring pixels, DCENet does not use Batch Normalization after the convolutional layers. The activation function for the hidden layers is ReLU, and the activation function for the output layer is tanh considering the output range is [-1, 1]. The 6 hidden layers utilize symmetric skip connections similar to UNet. Specifically, the outputs of layers 3 and 4 are concatenated together with skip connection and then fed into the fifth layer. The outputs of layers 2 and 5 are concatenated together and fed into the sixth layer. Finally, the outputs of layers 1 and 6 are concatenated and fed into the seventh layer.

### Zero-DCE++

Although Zero-DCE [15] is already a fast and concise network, improving efficiency remains significant, particularly when dealing with large images. Zero-DCE++ [1] focuses on reducing computational cost and achieving faster inference speed while maintaining performance.

The first refinement concerns the convolutional layers used in DCENet. Similar to MobileNet

[68], Zero-DCE++ [1] replaces the convolutional layers of DCENet with depthwise separable convolutions to reduce the network parameters. A depthwise separable convolutional layer consists of a depthwise convolution with 3x3 kernels and a stride of 1, followed by a pointwise convolution with 1x1 kernels and a stride of 1. Through research, it was found that there is not much change in the estimated curve parameters across the 8 iterations. Therefore, Zero-DCE++ reduces the number of output curve parameter maps from 24 to 3. As a result, Equation (3.6) can be simplified to:

$$LE_n(x) = LE_{n-1}(x) + A(x)I(x)(1 - LE_{n-1}(x)) \tag{3.7}$$

Where A(x) becomes stable in different iterations. Despite reusing the curve parameter maps, they maintain their high-order characteristics due to the iterative process.

Since Zero-DCE is not sensitive to input image size, downsampling the image initially is a good way to reduce convolutional overhead. I can downsample the input image first and upsample it to restore the original size later on. The enhancement degree of pixels in a local region, which can be represented by curve parameters, is similar. Therefore, downsampling and upsampling do not greatly affect the model's final performance. Furthermore, the loss function used to maintain spatial information also contributes to the restoration process.

**Loss Functions**

Zero-DCE [15] and Zero-DCE++ [1] utilize four loss functions to constrain the enhanced image from various perspectives without reference data. These non-reference losses are spatial consistency loss, exposure control loss, color constancy loss, and illumination smoothness loss. Let's introduce them separately below.

Spatial Consistency Loss. Denoted as $L_{spa}$, the spatial consistency loss function ensures the spatial information of the image by controlling the difference between the value of a pixel and its neighboring pixels, ensuring it does not change excessively. The equation is:

$$L_{spa} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j\in\Omega(i)}(|Y_i - Y_j| - |I_i - I_j|)^2 \tag{3.8}$$

Where K is the number of local regions and i represents the traversal of local regions. $\Omega(i)$ is the four neighboring regions of $i_{th}$ region. I represents the average intensity value of the local in region enhanced image, and Y is the same value in input image.

In actual implementation, I don't have to ensure that the difference between every pixel's value and its surrounding pixels' values remains unchanged. Instead, in reality, I use local pixel regions of size 4x4, denoted by i, where each region's value is the average of all pixel values within it. This can be calculated through average pooling.

Exposure Control Loss. Denoted as $L_{exp}$, the exposure control loss function controls the exposure level and adjusts the exposure value of each pixel closer to a certain intermediate value to restrain under-exposed and over-exposed regions. To achieve this goal, Zero-DCE sets a

well-exposedness level E and aims to make the average intensity value of a local region close to E, which means:

$$L_{exp} = \frac{1}{M}\sum_{k=1}^{M}(Y_k - E) \tag{3.9}$$

Where M is the number of non-overlapping local regions of size 16x16. Similar to the equation of spatial consistency loss, in this function, Y is the average intensity value of a non-overlapping local region in the enhanced image.

In actual settings, E is set to 0.6 for light enhancement. E can also be set to other values for different exposure conditions as I discussed in previous section.

Color Constancy Loss. Denoted as $L_{col}$, the color constancy loss function ensures that the values of a certain color channel in the image should not significantly exceed those of other channels. Based on the Gray-World color constancy hypothesis [69], the values of a certain color channel in the image should not significantly exceed those of other channels. Color constancy loss is designed to prevent significant differences in values between different color channels in the enhanced image. Additionally, the color constancy loss also establishes correlations between different color adjustments.

The equation of color constancy loss is shown below:

$$L_{col} = \sum_{\forall(p,q)\in\varepsilon}(J_p - J_q)^2, \varepsilon = \{(R,G),(R,B),(G,B)\} \tag{3.10}$$

Where $J_p$ represents the average brightness of color channel p and (p, q) iterates through all pairwise combinations of the three color channels.

Illumination Smoothness Loss. Denoted as $L_{tv_A}$, this loss function maintain the monotonic relationship between neighboring pixels. The approach is to make neighboring pixels get closer parameter $\alpha \in A$.

$$L_{tv_A} = \frac{1}{N}\sum_{n=1}^{N}\sum_{c\in\xi}(|\Delta_x A_n^c| + |\Delta_y A_n^c|)^2, \xi = \{R,G,B\} \tag{3.11}$$

where N is the number of iteration and $\Delta_x$ and $\Delta_y$ are the horizontal and vertical gradient operators respectively. For images, the horizontal gradient and vertical gradient are simply the differences between the values of adjacent pixels to the left and above, respectively.

The total loss is the weighted sum of the above four errors:

$$L_{up} = L_{spa} + W_1 L_{exp} + W_2 L_{col} + W_3 L_{tv_A} \tag{3.12}$$

We can get upstream loss reformed by the $L_{total}$, denoted as $L_{up}$. And I get the upstream loss and output image which is the input of the downstream model:

$$\{I_{output}, L_{up}\} = model_{up}(I_{Input}) \tag{3.13}$$

### 3.2.4 Upstream Model Refining

The downstream model, serving as a fixed pre-trained model, remains stable during the training process. My objective is to enhance the performance of the upstream model to align with the downstream task. To accomplish this, I iteratively update the upstream model by combining the upstream loss, computed from the input images, with the downstream loss propagated back from the downstream model.

$$L_{total} = W_{up}L_{up} + W_{down}L_{down} \tag{3.14}$$

To balance the impact of the upstream and downstream losses, I introduce the weights $W_{up}$ and $W_{down}$. The final loss value is then recalculated and propagated back to the upstream model for parameter updates. This integration enables the upstream model to learn from downstream tasks and adjust its parameters accordingly, thereby enhancing its ability to generate enhanced images that are better suited for downstream tasks.

# Chapter 4

# Test and Experiments

## 4.1 Dataset

I use the train and test data of WIDER FACE dataset [3] to generate low light images for the upcoming experiments. This dataset is a widely-used benchmark for face detection tasks, comprising 32,203 images with annotations for 393,703 faces. It presents various challenges including data volume, annotation complexity, scene diversity, facial scale, occlusion, and pose variations, making it suitable for evaluating the robustness and effectiveness of face detection algorithms.

The WIDER FACE dataset [3] is partitioned into training, validation, and testing sets, with a split ratio of 50% for training, 10% for validation, and 40% for testing. This partitioning ensures a balanced distribution across different event classes within each subset. Each subset is further categorized into three levels of detection difficulty: Easy, Medium, and Hard. The dataset contains faces with significant variations in scale, pose, lighting conditions, expression, and occlusion, making it highly representative of real-world scenarios. With its extensive data volume, diverse labels, and facial diversity, the WIDER FACE dataset [3] is widely utilized in face detection tasks.

## 4.2 Experimental Settings

For the downstream model, I import pretrained model weights since it does not require training. However, to obtain loss values for my experiments, I still simulate the training process without updating the model weights. The downstream model employs ResNet-50 [52] as its backbone, augmented with a 6-level Feature Pyramid Network (FPN) [46], to extract multi-scale features from the input image.

For the upstream model, I set the batch size to 4 and initialize the filter weights of each layer following a Gaussian distribution with a standard deviation of 0.02 and a mean of 0.

I opt for the ADAM optimizer for model updates, using default parameters, and set the learning rate to $1e^{-4}$ for model optimization. Regarding the weights for the upstream loss, I set $W_1$ to 10, $W_2$ to 5 and $W_3$ for 1600. For the $W_{up}$ and $W_{down}$, I choose a ratio of 2:1, which meaning

$W_{up}$ is set to 1 and $W_{down}$ is set to 0.5. I will explore the the values of $W_{up}$ and $W_{down}$ later.

Regarding the test data, I organize four experimental groups of images as follows to compare the average precision (AP) values for the subsequent face detection task.

1. Low light images.
2. Enhanced images by Zero-DCE model only.
3. Ehanced images by my joint training model.
4. original normal light images

Group 1 is the generated training low light images while group 2 is the original data, which can be seen as ground truth. Group 2 and 3 are enhanced images from the models with and without joint training framework.

## 4.3 Test Results

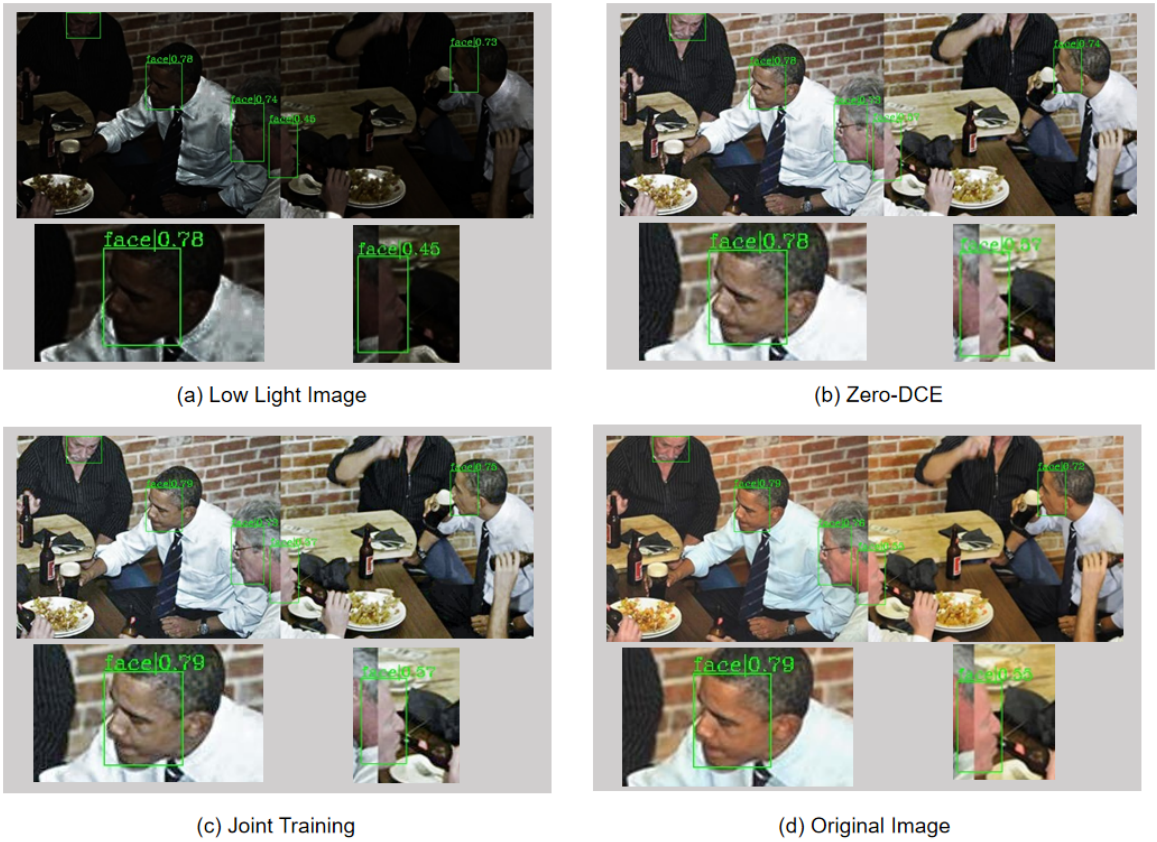
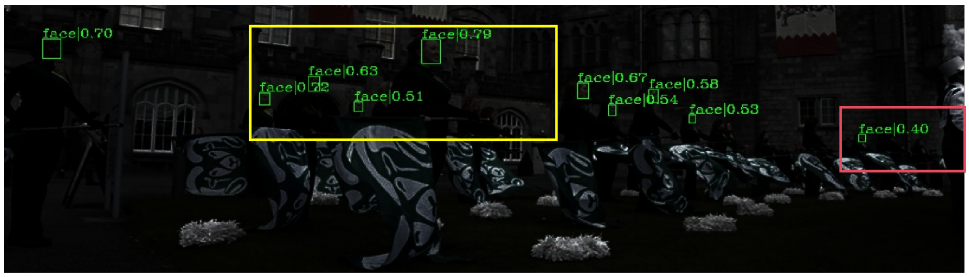


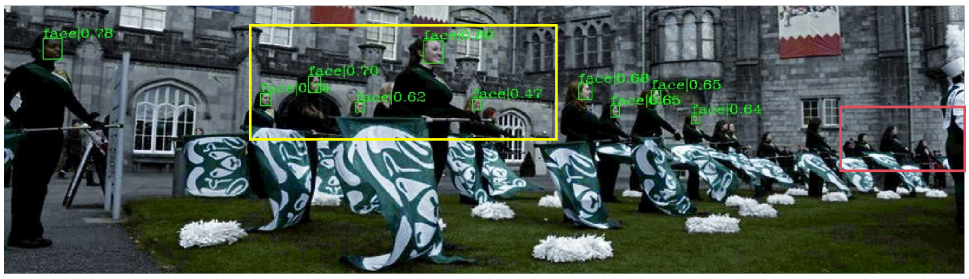

Figure 4.1: The test result of general faces.

I feed the four groups of images into the downstream face detection model and analyze the resulting detection outputs. In terms of general face detection, in theory, images enhanced by the joint training framework are expected to yield higher confidence values for the detected

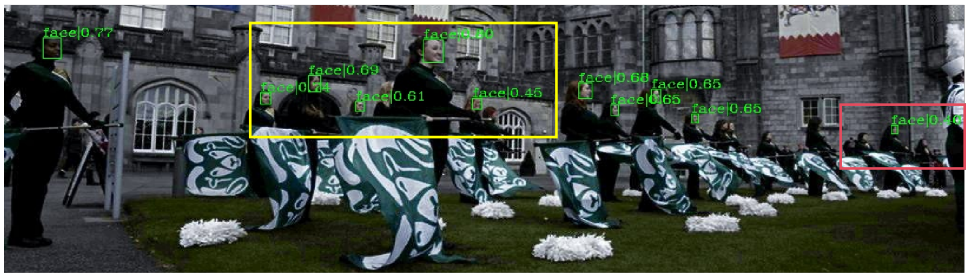
boxes. However, since Tinaface [2] is already a robust face detection model, it can still achieve satisfactory detection results even for regular-sized faces under low-light conditions.



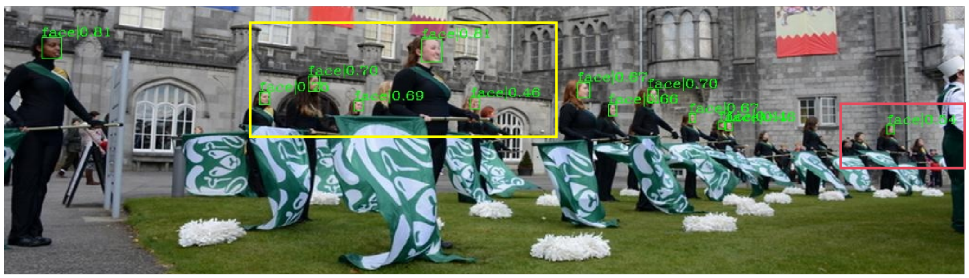

(a) Low Light Image




(b) Zero-DCE




(c) Joint Training




(d) Original Image

Figure 4.2: The test result of small faces.

As illustrated in Fig. 4.1, while low light does have an impact on the face detection task, as evidenced by the lower scores of the low-light images, the magnitude of score reduction is not significant. This makes it challenging for us to directly quantify the impact of different methods. However, fortunately, the enhanced images generated by my method still bear a closer resemblance to the recognition results of normal light images and ground truth. This confirms that joint training indeed assists the upstream model in understanding the requirements of the downstream model in certain scenarios.

Advanced experiments are required to compare the different methods thoroughly. Additionally,

I observe that low-light conditions pose a greater challenge for smaller face detection. Figure 4.1 also illustrates that low-light images have a more significant impact on smaller faces, and the degree of improvement in score for enhanced images on smaller faces is higher.

Small face detection poses a significant challenge in face detection because general faces contain enough information for face detection models to effectively process. However, smaller faces contain fewer pixels, which means there is less effective information available for detection. In low-light conditions, this problem is exacerbated as the reduced lighting further diminishes the available information for detecting small faces.

The issue is illustrated in Fig. 4.2. My focus is on the area within the yellow box. In Fig. 4.2(a), the downstream model fails to detect some faces in low-light images that are successfully recognized in Fig. 4.2(d). For the recognized faces, the confidence scores are much lower than those in Fig. 4.2(d) in normal light. However, in Fig. 4.2(b) and 4.2(c), the faces that were not detected in Fig. 4.2(a) are successfully recognized by the same downstream model. This demonstrates that the LLIE model not only enhances the visual effect of low-light images but also recovers important information for the downstream task.

However, although LLIE methods can recover lost information, the training process of the LLIE model has not considered the downstream model, so the LLIE model does not gain the right direction towards recovery. The refinement is not stable and obvious. I can observe this in the red box in Fig. 4.2. In Fig. 4.2(a), the downstream model successfully recognizes the face, although the score is lower. However, in Fig. 4.2(b), the enhanced image from Zero-DCE even loses information, causing the downstream model to fail to recognize the face. This demonstrates that a better visual effect does not necessarily translate to better performance in downstream tasks. When the LLIE model enhances the visual effect, guidance is needed to improve the results of the image in downstream tasks.

To solve this problem, my framework introduces the information backpropagated from the downstream model to guide the upstream model in enhancing the image while preserving and further recovering the information crucial for downstream tasks. In Fig. 4.2(c), the enhanced image from my framework enables the downstream model to recognize the lost face in Fig. 4.2(b) and achieve a higher score.

## 4.4 Evaluation Results

The evaluation experiment is applied to the four groups of images, and the Average Precision (AP) for face detection is calculated using the test set of the dark WIDERFACE dataset [3]. AP (Average Precision) is a commonly used metric for evaluating the performance of object detection models. It measures the average precision at different confidence thresholds. During the evaluation process, the model's predictions for detection boxes are typically sorted by confidence scores, and the precision and recall for each detection box are calculated. Then, the AP is computed from the precision-recall curve. The AP value ranges from 0 to 1, where higher values indicate better performance of the model in the detection task. Models with higher AP values are generally considered to have better performance.

The AP performance of the 4 groups of images is shown in Table 4.1. Compared with the results of the ground truth, which are 96.31%, 95.64%, and 93.01% in the three settings, low-light images indeed have a significant impact on the task of face recognition, with AP values

of 94.04%, 95.59% in the three settings, respectively. The impact is particularly severe for the Hard set. Since the data in the Hard set covers the Medium and Easy sets, the performance on the Hard set can better represent the overall results of the data.

| Methods | easy | medium | hard |
|---|---|---|---|
| Low light images | 0.9404 | 0.9259 | 0.8852 |
| Zero-DCE | 0.9437 | 0.9312 | 0.8904 |
| Joint training | 0.9478 | 0.9387 | 0.8998 |
| Original images | 0.9631 | 0.9564 | 0.9301 |

Table 4.1: AP performance of four groups of test images

The AP value of the enhanced images of Zero-DCE [15] shows improvement, demonstrating that low-light enhancement enhances the overall performance of the downstream model in face recognition. However, the improvement across the three different sets (Easy, Medium, and Hard) is only 0.33%, 0.53%, and 0.52%, respectively. In comparison, enhanced images from the joint training model show greater improvements, with performance increases of 0.74%, 1.28%, and 1.46% on the three settings, respectively. This suggests that joint training guides the upstream models to further improvement, thus better aligning with the face recognition task of downstream models.

Notably, compared with the original images, the degree of decrease in AP value of the low-light images is the highest on the Hard set. Similarly, the degree of increase in AP value of the enhanced images from my model is also the highest on the Hard set. The Hard set not only presents higher requirements for face detection tasks but also encompasses the content from the Easy and Medium sets. Therefore, the performance on the Hard set can better represent the overall performance of the model on unseen data. These comparative results further demonstrate that the joint training framework improves the low-light enhancement approach to be more aligned with the requirements of downstream tasks.

## 4.5 The Loss Weights

In my experimental settings, I chose the ratio of $W_{up}$ to $W_{down}$ as 2:1. Although the upstream loss $L_{up}$ and downstream loss $L_{down}$ have been transformed to the same range from the original loss values, I still need $W_{up}$ and $W_{down}$ to adjust their proportion in the final loss to better guide model training. These two weights respectively represent the impact of the upstream loss and downstream loss on the training of the upstream model.


Figure 4.3: The enhanced images with different $W_{up}$ and $W_{down}$ values.

To determine the best values of $W_{up}$ and $W_{down}$, I experimented with different ratios of $W_{up}$ to $W_{down}$ to obtain enhanced images. Then I compared their AP values in downstream tasks to identify which setting yielded the greatest improvement for downstream tasks.

| $W_{up} : W_{down}$ | easy | medium | hard |
|---|---|---|---|
| 1:0 | 0.9437 | 0.9312 | 0.8904 |
| 1:1 | 0.9464 | 0.9362 | 0.8974 |
| 3:2 | 0.9465 | 0.9363 | 0.9380 |
| 2:1 | 0.9478 | **0.9387** | **0.8998** |
| 3:1 | **0.9480** | 0.9375 | 0.8987 |
| 2:3 | 0.9454 | 0.9351 | 0.8967 |
| 1:3 | 0.9457 | 0.9352 | 0.8962 |
| 0:1 | 0.9418 | 0.9294 | 0.8889 |

Table 4.2: AP performance of different ratio of $W_{up}$ and $W_{down}$

The ratio of $W_{up}$ to $W_{down}$ was set as 2:1, which resulted in the best AP value on the medium and hard sets. Meanwhile, for the simple set, the ratio of 3:1 performed the best. Since I believe that the performance on the hard set can better reflect the effectiveness of the model in general applications, and the difference between the results of 2:1 and 3:1 ratios on the simple set was only 0.02%, I selected the 2:1 ratio as the optimal setting.

From Table 4.2, it is evident that the results obtained after introducing $loss_{down}$ are significantly better than using Zero-DCE alone, highlighting the importance of guidance from the downstream model. However, if I completely abandon $loss_{up}$, I obtain the worst result, only slightly better than the results from purely low-light images. Ultimately, the goal of the upstream model is to enhance low-light images. The improvement in downstream model results after Zero-DCE enhancement underscores that downstream loss is not a substitute for upstream loss.

The result of $2:1$ is better than that of $1:1$. I learn that the impact of upstream loss on the recovery of effective information is greater. The LLIE model itself has the ability to recover effective information. The role of downstream loss is to provide certain directional guidance during upstream model training, enabling better performance of results for downstream tasks. Therefore, upstream loss should have higher weights during training. However, since I have already demonstrated the effect of the downstream loss, if the weight of the downstream loss is too small, its guidance on the upstream model will also diminish, resulting in a smaller improvement margin.

# Chapter 5

# Conclusion and Recommendations

In contemporary times, the advancement of low-light image enhancement technology has emerged as a focal point in both academia and industry. By enhancing the brightness, contrast, and details of low-light images, this technology aims to render them more suitable for human observation and computer analysis, promising a plethora of application prospects.

While numerous LLIE methods exist to enhance low-light images, many overlook the importance of considering the impact on subsequent downstream vision tasks, which is the central focus of mt project. I propose a task-oriented joint training framework to integrate the LLIE model with the performance of enhanced images in downstream models. Following enhancement, the output image serves as input for the downstream model, and the loss values of the downstream model are fed back to the upstream model. These losses are merged with the upstream loss to guide refinement of the upstream model. Specifically, I utilize Zero-DCE++ as the upstream model and Tinaface as the downstream model to assess the effectiveness of my framework. Experimental results confirm that my framework indeed enhances the alignment of the upstream model with the requirements of downstream models. Additionally, I investigate the impact of different ratios of upstream and downstream losses on the results. Ultimately, I demonstrate the necessity of assigning greater weight to the upstream loss while acknowledging the indispensability of the downstream loss

However, I must acknowledge that the downstream loss, reliant on the execution of the downstream model, can diminish training efficiency. Additionally, for further advancements, I aim to improve the calculation and management of the downstream loss, facilitating better integration with the upstream loss. In subsequent experiments, we plan to explore additional downstream tasks in various fields. These may include object detection and semantic segmentation. Furthermore, I will assess the effectiveness of alternative LLIE methods.

# Bibliography


[1] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(8):4225–4238, 2021.

[2] Yanjia Zhu, Hongxiang Cai, Shuhan Zhang, Chenhao Wang, and Yichao Xiong. Tinaface: Strong but simple baseline for face detection. arXiv preprint arXiv:2011.13183, 2020.

[3] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5525–5533, 2016.

[4] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In European conference on computer vision, pages 573–589. Springer, 2022.

[5] Haidi Ibrahim and Nicholas Sia Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. IEEE Transactions on Consumer Electronics, 53(4):1752–1758, 2007.

[6] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Transactions on image processing, 16(8):2080–2095, 2007.

[7] Vijay A Kotkar and Sanjay S Gharde. Review of various image contrast enhancement techniques. International journal of innovative research in Science, Engineering and Technology, 2(7), 2013.

[8] Edwin H Land. An alternative technique for the computation of the designator in the retinex theory of color vision. Proceedings of the national academy of sciences, 83(10):3078–3080, 1986.

[9] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560, 2018.

[10] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. IEEE transactions on image processing, 22(9):3538–3548, 2013.

[11] Xueyang Fu, Yinghao Liao, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A probabilistic method for image enhancement with simultaneous illumination and reflectance estimation. IEEE Transactions on Image Processing, 24(12):4965–4977, 2015.

[12] Xutong Ren, Wenhan Yang, Wen-Huang Cheng, and Jiaying Liu. Lr3m: Robust low-light enhancement via low-rank regularized retinex model. IEEE Transactions on Image Processing, 29:5862–5876, 2020.

[13] Seonhee Park, Soohwan Yu, Byeongho Moon, Seungyong Ko, and Joonki Paik. Low-light image enhancement using variational optimization-based retinex model. IEEE Transactions on Consumer Electronics, 63(2):178–184, 2017.

[14] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. IEEE transactions on image processing, 30:2340–2349, 2021.

[15] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1780–1789, 2020.

[16] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE Transactions on Image processing, 6(7):965–976, 1997.

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

[18] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In BMVC, volume 220, page 4. Northumbria University, 2018.

[19] Feifan Lv, Yu Li, and Feng Lu. Attention guided low-light image enhancement with a large scale low-light simulation dataset. International Journal of Computer Vision, 129(7):2175–2193, 2021.

[20] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16, pages 492–511. Springer, 2020.

[21] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2881–2890, 2017.

[22] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In Proceedings of the IEEE/CVF international conference on computer vision, pages 2482–2491, 2019.

[23] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In Proceedings of the 27th ACM international conference on multimedia, pages 1632–1640, 2019.

[24] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. International Journal of Computer Vision, 129:1013–1037, 2021.

[25] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. Low-light image enhancement via a deep hybrid network. IEEE Transactions on Image Processing, 28(9):4364–4375, 2019.

[26] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5901–5910, 2022.

[27] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Low-light image and video enhancement using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(12):9396–9416, 2021.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015.

[29] Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 12918–12927, 2023.

[30] Khurram Azeem Hashmi, Goutham Kallempudi, Didier Stricker, and Muhammad Zeshan Afzal. Featenhancer: Enhancing hierarchical features for object detection and beyond under low-light vision. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6725–6735, 2023.

[31] Jinxiu Liang, Jingwen Wang, Yuhui Quan, Tianyi Chen, Jiaying Liu, Haibin Ling, and Yong Xu. Recurrent exposure generation for low-light face detection. IEEE Transactions on Multimedia, 24:1609–1621, 2021.

[32] Wenjing Wang, Xinhao Wang, Wenhan Yang, and Jiaying Liu. Unsupervised face detection in the dark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1):1250–1266, 2022.

[33] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. Computer Vision and Image Understanding, 178:30–42, 2019.

[34] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. IEEE Transactions on Image Processing, 29:5737–5752, 2020.

[35] Bo Zhang, Yuchen Guo, Runzhao Yang, Zhihong Zhang, Jiayi Xie, Jinli Suo, and Qionghai Dai. Darkvision: a benchmark for low-light image/video perception. arXiv preprint arXiv:2301.06269, 2023.

[36] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 1792–1800, 2022.

[37] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In Proceedings of the IEEE/CVF international conference on computer vision, pages 2553–2562, 2021.

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.

[39] Xinwei Xue, Jia He, Long Ma, Yi Wang, Xin Fan, and Risheng Liu. Best of both worlds: See and understand clearly in the dark. In Proceedings of the 30th ACM International Conference on Multimedia, pages 2154–2162, 2022.

[40] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. IEEE transactions on pattern analysis and machine intelligence, 43(2):652–662, 2019.

[41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 2012.

[42] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014.

[43] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015.

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 2015.

[45] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.

[46] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017.

[47] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pages 21–37. Springer, 2016.

[48] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

[49] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7263–7271, 2017.

[50] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[51] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015.

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

[53] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[55] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3588–3597, 2018.

[56] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In European conference on computer vision, pages 213–229. Springer, 2020.

[57] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In Proceedings of the IEEE international conference on computer vision, pages 192–201, 2017.

[58] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5060–5069, 2019.

[59] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017.

[60] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641, 2019.

[61] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. Advances in neural information processing systems, 29, 2016.

[62] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pages 694–711. Springer, 2016.

[63] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, pages 448–456. pmlr, 2015.

[64] Yuxin Wu and Kaiming He. Group normalization. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018.

[65] Shengkai Wu, Xiaoping Li, and Xinggang Wang. Iou-aware single-stage object detector for accurate localization. Image and Vision Computing, 97:103911, 2020.

[66] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 764–773, 2017.

[67] Yundong Zhang, Xiang Xu, and Xiaotao Liu. Robust and high performance face detector. arXiv preprint arXiv:1901.02350, 2019.

[68] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[69] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. In Computer graphics forum, volume 28, pages 161–171. Wiley Online Library, 2009.