A MAJOR PROJECT – II REPORT
ON
LIGHTWEIGHT NETWORKS FOR LOW-LIGHT IMAGE ENHANCEMENT BASED ON MULTI-OBJECTIVE KNOWLEDGE DISTILLATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE

OF

MASTER OF TECHNOLOGY
IN
COMPUTER SCIENCE & ENGINEERING

Submitted By
JOSHI KULADIPKUMAR NARSHIHBHAI
23/CSE/29

Under The Supervision Of
PROF. ARUNA BHAT
(Professor)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

May 2025

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Joshi Kuladipkumar Narshihbhai (23/CSE/29)**, hereby certify that the work which is being presented in the major project report II entitled **"Lightweight Networks for Low-Light Image Enhancement based on Multi-Objective Knowledge Distillation"** in partial fulfillment of the requirements for the award of the Degree of Master of Technology, submitted in the **Department of Computer Science and Engineering**, Delhi Technological University is an authentic record of my own work carried out during the period from [Start Date, e.g., August 2024] to May 2025 under the supervision of **Prof. Aruna Bhat**.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor (s)

Signature of External Examiner

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project titled **"Lightweight Networks for Low-Light Image Enhancement based on Multi-Objective Knowledge Distillation"**, submitted by **Joshi Kuladipkumar Narshihbhai**, Roll No. **23/CSE/29**, Department of **Computer Science & Engineering**, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of **Master of Technology (M.Tech)** in Computer Science and Engineering is a genuine record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree to this University or elsewhere.

Place: Delhi

Date: _____

Prof. Aruna Bhat

Professor

Delhi Technological University

ACKNOWLEDGEMENT

I am grateful to **Prof. Manoj Kumar, HOD** (Department of Computer Science and Engineering), Delhi Technological University (Formerly Delhi College of Engineering), New Delhi, and all other faculty members of our department for their astute guidance, constant encouragement, and sincere support for this project work.

I am writing to express our profound gratitude and deep regard to my project mentor **Prof. Aruna Bhat**, for her exemplary guidance, valuable feedback, and constant encouragement throughout the project. Her valuable suggestions were of immense help throughout the project work. Her perspective criticism kept us working to make this project much better. Working under her was an extremely knowledgeable experience for us.

I would also like to thank all my friends for their help and support sincerely.

Joshi Kuladipkumar Narshihbhai
(23/CSE/29)

ABSTRACT

Low-light image enhancement is a critical computer vision challenge impacting visual quality and downstream tasks. While large deep learning models like MIRNet-v2 excel at restoring details and colors, their computational and memory demands hinder real-time or edge device deployment. Conversely, lightweight models such as Zero-DCE efficiently adjust illumination but may lack fine texture or structural recovery.

This thesis introduces a novel framework for computationally efficient, high-performing lightweight networks for low-light image enhancement using progressive multi-teacher knowledge distillation. We leverage the complementary expertise of MIRNet-v2 (˜26.6M parameters) for structural preservation and Zero-DCE (˜0.1M parameters) for efficient illumination correction, training a compact student network to learn from both.

Key innovations include:

1. **Resolution-Progressive Training:** The student network trains over 250 epochs, starting with 64x64 pixel images and progressively increasing resolution (to 128x128, then 256x256). This curriculum learning, with dynamic batch sizing, ensures stable optimization and manages GPU memory.

2. **Multi-Objective Hybrid Loss:** A weighted 7-component loss guides student learning, incorporating Charbonnier loss ($\mathcal{L}_{\text{recon}}$) for reconstruction, FFT-based L1 loss ($\mathcal{L}_{\text{freq}}$) for frequency fidelity, VGG perceptual loss ($\mathcal{L}_{\text{perc}}$), MS-SSIM loss ($\mathcal{L}_{\text{msssim}}$) for structure, Sobel gradient L1 loss ($\mathcal{L}_{\text{grad}}$) for edges, a PatchGAN adversarial loss ($\mathcal{L}_{\text{adv-G}}$), and contrastive distillation loss ($\mathcal{L}_{\text{cont}}$) to mimic teacher features.

3. **Lightweight Enhanced Student Architecture:** The 9.99M parameter student generator uses an efficient backbone of Recursive Residual Groups (RRGs) and Multi-scale Residual Blocks (MRBs). Optional Adaptive Feature Stretch (AFS) blocks for dynamic feature range expansion and Gradient-Guided Convolution (GGC) blocks for edge awareness further augment its representational power.

Experiments on the standard LOL (Low-Light) dataset show our model achieving a PSNR of 21.89 dB and an SSIM of 0.858. With total training parameters (including discriminator) of 23.6M, our approach significantly balances performance and computational efficiency, offering a promising solution for deploying high-quality low-light enhancement in practical, resource-aware scenarios.

Contents

List of Tables

List of Figures

List of Symbols, Abbreviations, and Nomenclature

AFS	Adaptive Feature Stretch
AMP	Automatic Mixed Precision
CA	Channel Attention
CNN	Convolutional Neural Network
CV	Computer Vision
D	Discriminator (in GAN)
DL	Deep Learning
FFT	Fast Fourier Transform
G	Generator (in GAN)
GAN	Generative Adversarial Network
GAP	Global Average Pooling
GGC	Gradient-Guided Convolution
GPU	Graphics Processing Unit
HE	Histogram Equalization
KD	Knowledge Distillation
LOL	Low-Light (dataset name)
LR	Learning Rate
MIRNet	Multi-scale Interactive Residual Network
MRB	Multi-scale Residual Block
MS-SSIM	Multi-Scale Structural Similarity Index Measure
PSNR	Peak Signal-to-Noise Ratio
RRG	Recursive Residual Group
SSIM	Structural Similarity Index Measure
SOTA	State-Of-The-Art
Zero-DCE	Zero-Reference Deep Curve Estimation
S_{low}	Input Low-Light Image
S_{norm}	Ground Truth Normal-Light Image
$\hat{S}$	Enhanced Output Image from Student
G	Student Generator Network
T_1, T_2	Teacher Networks
F_{in}, F_{out}	Student Features (Input/Output of a stage)
F_s, F_t	Student and Teacher Features for Distillation
$\mathcal{L}^G_{\text{total}}$	Total Loss for Generator

Chapter 1
Introduction

1.1 Background and Motivation

The ability to capture and interpret visual information effectively is paramount in an increasingly digital world. However, images acquired under sub-optimal low-light conditions often suffer from a cascade of degradations. These include poor visibility due to insufficient photons, the prevalence of sensor noise (e.g., Poisson-Gaussian noise) which becomes more apparent with signal amplification, color distortions resulting from inaccurate white balancing or sensor characteristics under low illumination, and a significant loss of fine-grained details and textures that are crucial for both human perception and machine understanding [1]. Such degradations not only diminish the aesthetic appeal of photographs and videos but also severely compromise the performance of downstream computer vision applications. For instance, in autonomous navigation, poor visibility at night can lead to object detection failures; in surveillance, noisy low-light footage can hinder identification; and in medical imaging, subtle diagnostic features might be obscured, impacting diagnostic accuracy. The challenge, therefore, is not merely aesthetic but has profound practical implications across various domains.

Traditional image processing techniques, such as global Histogram Equalization (HE) [2] or its adaptive variants like Contrast Limited Adaptive Histogram Equalization (CLAHE) [3], attempt to improve contrast by redistributing pixel intensities. While simple and computationally inexpensive, these methods can lead to over-enhancement, noise amplification, and often produce unnatural-looking images, especially in scenes with heterogeneous lighting. Retinex theory [4], which models an image S as the product of scene reflectance R and illumination L (i.e., $S(x, y) = R(x, y) \cdot L(x, y)$), has inspired numerous enhancement algorithms. These methods typically aim to estimate the illumination component and then either modulate it or use it to derive the reflectance component, which is assumed to be invariant to lighting conditions [5]. However, the accurate and robust decomposition of an image into these two components is an inherently ill-posed problem. This often leads to challenges such as halo artifacts around strong edges, unnatural color rendition, and incomplete noise removal, particularly in scenes characterized by complex lighting and diverse content.

1.2 Problem Statement

The advent of deep learning has brought significant breakthroughs in low-light image enhancement. State-of-the-art (SOTA) models, exemplified by complex architectures like MIRNet-v2 [6], leverage deep and intricate neural networks. These models often employ sophisticated techniques such as multi-scale processing, attention mechanisms, and advanced residual learning strategies to effectively learn mappings from degraded low-light inputs to high-quality, well-exposed outputs. While they achieve remarkable performance in terms of objective metrics (e.g., Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM)) and subjective visual quality, their primary drawback lies in their substantial computational footprint. For instance, MIRNet-v2, in its default configuration for enhancement tasks, comprises approximately 26.6 million parameters. This large model size translates directly to high memory requirements and significant inference latency, rendering such models unsuitable for deployment on resource-constrained platforms. These platforms include mobile devices, embedded systems, or edge computing nodes, where real-time processing is often a critical requirement for practical applications.

Conversely, lightweight models such as Zero-Reference Deep Curve Estimation (Zero-DCE) [1] have emerged, placing a strong emphasis on computational efficiency. Zero-DCE, with fewer than 0.1 million parameters, learns image-adaptive tonal curves to adjust illumination without the need for paired training data. While highly efficient, its primary focus on illumination adjustment means it may not fully address other common low-light degradations, such as severe noise or the recovery of fine textural details, as effectively as larger, more complex models. This disparity highlights a critical performance-efficiency trade-off: achieving high-fidelity enhancement typically comes at the cost of high computational demand, whereas efficient models may compromise on the achievable output quality. There is, therefore, a pressing need for innovative techniques that can develop lightweight networks capable of delivering high-quality low-light image enhancement, effectively bridging this gap and enabling advanced image restoration on a wider range of devices and applications.

1.3 Objectives of the Thesis

The primary objective of this thesis is to develop and evaluate a lightweight deep learning model for low-light image enhancement that achieves a competitive balance between image quality and computational efficiency. This overarching goal is broken down into the following specific objectives:

1. To design an efficient student network architecture that is significantly smaller

in terms of parameter count compared to SOTA high-performance models like MIRNet-v2, yet remains capable of learning complex enhancement transformations effectively.

2. To investigate and implement a multi-teacher knowledge distillation strategy, strategically leveraging the complementary strengths of a structurally-focused teacher model (MIRNet-v2) and an illumination-focused teacher model (Zero-DCE) to comprehensively guide the student network's learning process.

3. To develop and apply a resolution-progressive training methodology. This involves starting the training with low-resolution images and gradually increasing to the target resolution, combined with dynamic batch size adjustment, to promote stable and effective training of the student model over a fixed number of epochs (e.g., 250 epochs).

4. To formulate and utilize a comprehensive hybrid loss function that incorporates multiple objectives. These include pixel-level fidelity, perceptual similarity, structural integrity, frequency-domain characteristics, gradient preservation, adversarial realism, and feature-level mimicry of teacher representations through contrastive distillation.

5. To empirically evaluate the proposed lightweight model on a standard low-light image enhancement dataset (e.g., the LOL dataset), comparing its performance in terms of PSNR and SSIM, as well as its parameter count, against the teacher models and other relevant benchmarks.

6. To analyze the impact of different components within the proposed framework, such as the progressive training strategy and key elements of the hybrid loss function. This analysis will be further substantiated by planned ablation studies in future work.

1.4 Scope of the Work

This research focuses on single-image low-light enhancement using supervised deep learning principles and knowledge distillation techniques. The scope of the work encompasses the following:

- Utilization of existing pre-trained models, specifically MIRNet-v2 and Zero-DCE, as fixed teacher networks. The internal workings and parameters of these teacher models are not modified during the student training process.

- The design and implementation of a novel student Convolutional Neural Network (CNN) architecture that incorporates computationally efficient

building blocks (e.g., Recursive Residual Groups (RRGs), Multi-scale Residual Blocks (MRBs)) and optional enhancement modules (e.g., Adaptive Feature Stretch (AFS), Gradient-Guided Convolution (GGC)).

- The development of a comprehensive training pipeline. This pipeline includes the resolution-progressive training strategy, dynamic batch sizing to manage GPU memory, and a multi-component hybrid loss function, with the training executed for a total of 250 epochs.
- Evaluation of the proposed model primarily on the LOL (Low-Light) paired dataset, which serves as a standard benchmark for this image enhancement task.
- Quantitative assessment using established image quality metrics, namely PSNR and SSIM, complemented by qualitative visual assessment of the enhanced images to judge perceptual quality.
- The primary programming environment for implementation and experimentation is Python, utilizing the PyTorch deep learning framework.

This work does not delve into unsupervised or zero-reference training paradigms for the student model itself (beyond leveraging Zero-DCE as one of the teachers). Furthermore, it does not explicitly address the enhancement of low-light video sequences or real-time hardware deployment considerations beyond reporting parameter counts and discussing conceptual efficiency. The focus remains on achieving a balance between enhancement quality and model compactness for single images.

1.5 Thesis Organization

This thesis is organized into five chapters, structured as follows:

- **Chapter 1: Introduction** provides the foundational context for the research, outlining the background and motivation, defining the problem statement, specifying the research objectives, and detailing the scope of the work. This chapter also includes the overall organization of the thesis.
- **Chapter 2: Literature Survey** presents a comprehensive review of existing work relevant to this thesis. This includes a survey of classical and deep learning-based methods for low-light image enhancement, an overview of various knowledge distillation techniques (such as single-teacher, multi-teacher, feature-based, and contrastive methods), and a discussion of progressive and curriculum learning strategies pertinent to the proposed training methodology.

- **Chapter 3: Proposed Methodology** details the novel framework developed in this research. This chapter describes the architectures of the teacher and student networks, the design principles behind the efficient enhancement blocks (AFS, GGC), the specifics of the resolution-progressive training strategy, and the formulation of the comprehensive multi-component hybrid loss function.

- **Chapter 4: Results and Analysis** presents the experimental setup, the datasets utilized for training and evaluation, and the metrics employed for performance assessment. It includes quantitative results comparing the proposed model with baseline methods, qualitative visual comparisons on diverse image samples, and an analysis of the training dynamics observed during the learning process. A discussion on planned ablation studies and current limitations is also included.

- **Chapter 5: Conclusion and Future Scope** summarizes the key contributions and findings of the thesis. It discusses the significance of the research outcomes and suggests potential avenues for future work and further improvements in the domain of efficient low-light image enhancement.

Following these main chapters, the thesis includes Appendices containing supplementary materials such as the List of Publications resulting from this research and copies of relevant conference acceptance correspondences. The Bibliography section lists all the cited references.

Chapter 2
Literature Survey

The challenge of effectively enhancing images captured under low-illumination conditions has spurred extensive research over several decades. This chapter provides a comprehensive review of existing methodologies, categorized into classical image processing techniques, modern deep learning-based approaches, relevant knowledge distillation strategies, and the principles of progressive and curriculum learning that inform our training methodology.

2.1 Low-Light Image Enhancement Techniques

Approaches to low-light image enhancement can be broadly classified into traditional signal processing methods and contemporary data-driven deep learning techniques.

2.1.1 Classical Methods

Prior to the dominance of deep learning, several image processing techniques were developed to tackle low-light conditions.

Histogram Equalization and Variants

Histogram Equalization (HE) is one of the earliest and simplest techniques for contrast enhancement. It operates by redistributing the pixel intensity values to achieve a more uniform histogram, thereby stretching the contrast across the entire dynamic range [7]. While global HE can improve overall contrast, it often leads to a washed-out appearance and can significantly amplify existing noise, particularly in regions with low variance.

To address these limitations, adaptive variants were proposed. Adaptive Histogram Equalization (AHE) applies HE to contextual regions of the image rather than globally, preserving local details better. However, AHE can still over-amplify noise in relatively homogeneous regions. Contrast Limited Adaptive Histogram Equalization (CLAHE) [3, 2] mitigates this by clipping the histogram at a predefined value before computing the cumulative distribution function, thus limiting the amplification factor. Despite these improvements, HE-based methods often

struggle to produce natural-looking results and may not effectively handle severe noise or color distortions common in low-light imagery.

Retinex-Based Methods

Retinex theory, introduced by Land [4], provides a model for human color perception, suggesting that an observed image S can be decomposed into a product of scene reflectance R and illumination L:

$$S(x, y) = R(x, y) \cdot L(x, y) \tag{2.1}$$

Reflectance R is considered an intrinsic property of the objects in the scene, representing the actual colors and details, while illumination L represents the lighting conditions. The goal of Retinex-based enhancement is typically to estimate and remove or adjust the illumination component L to recover the reflectance R.

Early algorithms included Single-Scale Retinex (SSR) [8], which estimates illumination by convolving the image with a Gaussian filter. Multi-Scale Retinex (MSR) [5] improved upon SSR by combining the outputs of several SSR filters with different scales, offering better dynamic range compression and detail rendition. To address color distortions often introduced by MSR, the Multi-Scale Retinex with Color Restoration (MSRCR) algorithm was proposed, incorporating color constancy adjustments [9].

While Retinex-based methods offer a physically grounded approach, they face significant challenges. The decomposition into reflectance and illumination is an ill-posed problem, and inaccurate estimations can lead to halo artifacts around strong edges, unnatural color shifts, and insufficient noise suppression. Furthermore, many classical Retinex methods rely on hand-crafted parameters that may not generalize well across diverse scenes.

Dehazing-Inspired Techniques

Interestingly, techniques developed for image dehazing have found some application in low-light enhancement due to analogies between the scattering medium in haze and the underexposure/noise in low-light images. The Dark Channel Prior (DCP) [10], a popular dehazing prior, assumes that in most non-sky local patches, at least one color channel has some pixels with very low intensity. While not directly applicable, the underlying principle of identifying and manipulating specific image statistics has inspired some enhancement approaches that aim to ”invert” the effects of low light [11]. However, the physical models for haze and low-light noise are distinct, limiting the direct applicability and performance of such methods.

Frequency Domain Methods

Techniques operating in the frequency domain, such as those based on wavelet transforms or Fourier transforms, have also been explored. These methods aim to separate image components (e.g., illumination and detail, or signal and noise) in different frequency sub-bands, process them independently, and then reconstruct the enhanced image. For example, illumination can be associated with low-frequency components and details with high-frequency components. While offering potential for targeted noise reduction and detail enhancement, frequency domain methods can introduce ringing artifacts or other visual distortions if not carefully designed [12].

2.1.2 Deep Learning-Based Methods

The advent of deep convolutional neural networks (CNNs) has led to a paradigm shift in low-light image enhancement, with data-driven approaches consistently outperforming classical methods in terms of both objective metrics and subjective visual quality.

Early Convolutional Neural Network Approaches

Early deep learning models for low-light enhancement often focused on learning a direct end-to-end mapping from low-light images to their corresponding normal-light counterparts. LLNet [13] was a pioneering work that utilized a stacked sparse denoising autoencoder architecture. Other approaches employed various CNN architectures, often inspired by successful models in other image restoration tasks like denoising or super-resolution [14]. While demonstrating the potential of deep learning, these initial models sometimes struggled with generalization to unseen lighting conditions or produced results with residual noise or artifacts. The requirement for large datasets of paired low-light/normal-light images was also a significant challenge.

Retinex-Inspired Neural Networks

To incorporate physical priors and improve interpretability, several researchers integrated Retinex theory into deep learning frameworks. RetinexNet [15] proposed a two-stage network: a Decom-Net to decompose the input image into reflectance and illumination maps, and an Enhance-Net to adjust the illumination map and remove noise from the reflectance map. The final enhanced image is then reconstructed. This approach allows for more targeted manipulation of image components. KinD (Kindling the Darkness) [16] and its successor KinD++ [17] further refined this idea with more sophisticated network designs for decom-

position and enhancement, incorporating losses that explicitly considered image quality aspects like illumination smoothness and reflectance consistency. While these methods offer better control and physical grounding, their performance is highly dependent on the accuracy of the learned decomposition, and errors in one stage can propagate to the next.

Zero-Reference and Unsupervised Learning

Collecting large-scale datasets of perfectly aligned low-light and normal-light image pairs can be challenging and expensive. This has motivated the development of zero-reference or unsupervised learning techniques. Zero-DCE (Zero-Reference Deep Curve Estimation) [1], one of our teacher models, is a prime example. It trains a lightweight CNN to estimate pixel-wise higher-order curves for dynamic range adjustment. Crucially, Zero-DCE does not require paired data; instead, it is trained using a set of carefully designed non-reference loss functions, including spatial consistency loss, exposure control loss, color constancy loss, and illumination smoothness loss. This allows it to be trained on diverse low-light images without corresponding ground truths. EnlightenGAN [18] employs an unpaired Generative Adversarial Network (GAN) approach, using a global-local discriminator structure and self-regularized perceptual loss. While these methods offer significant advantages in terms of data requirements and efficiency, they might not always achieve the same level of fine detail recovery or noise suppression as SOTA supervised methods, as the learning signal is derived indirectly from image properties rather than direct comparison to a clean target.

State-of-the-Art High-Performance Models

Recent years have seen the emergence of powerful, often large-scale, architectures that have set new benchmarks in various image restoration tasks, including low-light enhancement. MIRNet [19] and its successor MIRNet-v2 [6] (our primary teacher model) are notable examples. MIRNet-v2 introduces a multi-scale residual block (MRB) that maintains high-resolution feature representations throughout the network via parallel convolutional streams operating at different spatial scales, while also facilitating information exchange across these streams. This design helps in preserving fine details while aggregating rich contextual information. The architecture is typically built upon a series of Recursive Residual Groups (RRGs), each containing multiple MRBs.

Restormer [20] brought Transformers [21] to the forefront of image restoration, proposing an efficient Transformer variant with multi-Dconv head transposed attention and gated-Dconv feed-forward networks, demonstrating strong performance on tasks including low-light enhancement. Other architectures like

HWMNet [22] have explored hierarchical wavelet-based multi-scale networks. While these models achieve excellent visual quality and objective scores, their primary limitation is their computational complexity and large number of parameters (e.g., MIRNet-v2 default config has ˜26.6M parameters), which makes them challenging for real-time applications.

Lightweight Network Designs for Efficiency

The demand for on-device AI and real-time image processing has spurred research into lightweight network architectures. Beyond Zero-DCE, general principles for creating efficient CNNs include using depthwise separable convolutions (as in MobileNets [23]), group convolutions, channel shuffling (ShuffleNets [24]), and network pruning and quantization techniques. While not always directly targeted at low-light enhancement, these architectural motifs provide inspiration for designing compact student networks. Our student architecture incorporates efficient blocks like RRGs and MRBs, which themselves can be made more lightweight by adjusting feature dimensions and using group convolutions.

Our work seeks to distill the strong restoration capabilities of a model like MIRNet-v2 and the illumination adjustment proficiency of Zero-DCE into a student network that is significantly more lightweight than the former, yet more powerful than the latter.

2.2 Knowledge Distillation Techniques

Knowledge Distillation (KD) provides a framework for transferring the "knowledge" from a large, cumbersome teacher model (or an ensemble of teachers) to a smaller, more efficient student model, aiming for the student to achieve performance comparable to the teacher but with reduced computational cost [25]. This is particularly relevant for deploying complex models on resource-constrained devices.

2.2.1 Response-Based Knowledge Distillation

The seminal work by Hinton et al. [25] focused on classification tasks, where the student network was trained to match the softened probability distribution (logits passed through a softmax with a temperature parameter $T > 1$) produced by the teacher network. The loss function typically combines a standard cross-entropy loss with the ground truth labels and a distillation loss term (e.g., KL divergence) that penalizes differences between the student's and teacher's softened outputs. This "dark knowledge" captured in the relative probabilities of incorrect classes was found to be a rich source of information for the student.

2.2.2 Feature-Based Knowledge Distillation

Instead of (or in addition to) matching outputs, feature-based KD aims to make the student's intermediate feature representations similar to those of the teacher. FitNets [26] was an early proponent, training thinner but deeper student networks to regress the feature maps from wider, pre-trained teacher networks at specific "hint" layers. Various strategies exist for matching features:

- **Direct L1/L2 Matching:** Minimizing the L1 or L2 distance between student and teacher feature maps after appropriate adaptation (e.g., using a convolutional regressor if channel dimensions differ).
- **Attention Transfer (AT):** Zagoruyko and Komodakis [27] proposed transferring attention by matching spatial attention maps derived from the sum of absolute values or squared values of feature activations across channels.
- **Matching Feature Statistics:** Some methods focus on matching statistical properties of feature maps, such as their mean and variance [28], or Gram matrices which capture feature correlations [29].

Feature-based distillation is particularly relevant for low-level vision tasks where intermediate features encode rich structural and textural information.

2.2.3 Relation-Based Knowledge Distillation

Relation-based KD shifts the focus from matching absolute values of individual features or outputs to matching the relationships between them. Park et al. [30] proposed distilling inter-sample relationships, such as the distance or angle between feature embeddings of different input samples. This encourages the student to learn a similarly structured feature space as the teacher.

2.2.4 Multi-Teacher Knowledge Distillation

When multiple teacher models are available, each potentially excelling in different aspects of a task or possessing diverse knowledge, a student can benefit from learning from all of them. Strategies for multi-teacher KD include averaging teacher outputs [31], using attention mechanisms to weight teacher contributions, or even training teachers adversarially. Our framework leverages two distinct teachers: MIRNet-v2 for structural fidelity and Zero-DCE for illumination expertise. While Zero-DCE's guidance is more implicit (through the student's efforts to produce well-illuminated images that satisfy various loss terms), MIRNet-v2's features are directly used in our contrastive distillation component.

2.2.5 Contrastive Representation Distillation (CRD)

CRD, proposed by Tian et al. [32], applies principles from contrastive self-supervised learning to the distillation problem. The core idea is to train the student so that its feature representation for a given input is close to the teacher's representation for the same input (positive pair) while being far from the teacher's representations for other inputs in the batch (negative pairs). This is typically achieved using an InfoNCE-style loss [33]:

$$\mathcal{L}_{\text{cont}} = -\sum_{i} \log \frac{\exp(sim(f_s(x_i), f_t(x_i))/\tau)}{\sum_j \exp(sim(f_s(x_i), f_t(x_j))/\tau)} \tag{2.2}$$

where f_s and f_t are student and teacher feature extractors (or projection heads), $sim(\cdot, \cdot)$ is a similarity function (e.g., cosine similarity), and τ is a temperature parameter. CRD has shown strong performance by encouraging the student to capture more fine-grained similarities in the learned feature space. Our $\mathcal{L}_{\text{cont}}$ term is based on this principle.

2.2.6 Knowledge Distillation in Low-Level Vision Tasks

KD has found successful applications in various low-level vision tasks. For instance, in image super-resolution, student networks learn from larger SR teachers to achieve good perceptual quality with fewer parameters [34]. In image denoising, KD can help lightweight networks learn effective noise suppression strategies [35]. Our work applies multi-faceted KD to the complex task of low-light image enhancement.

2.3 Progressive and Curriculum Learning Strategies

The idea of training machine learning models by starting with simpler concepts or data and gradually increasing complexity has strong parallels with human learning and can lead to improved training outcomes.

2.3.1 Curriculum Learning Fundamentals

Bengio et al. [36] formally introduced Curriculum Learning (CL), proposing that organizing training examples in a meaningful order (from easy to hard) can guide optimization towards better local minima and improve generalization. The definition of "easy" and "hard" can vary: it could be based on data characteristics (e.g., less noise, smaller objects), model capacity (starting with a simpler model), or task complexity. CL can be seen as a form of continuation method, helping to navigate non-convex optimization landscapes.

2.3.2 Progressive Growing in Generative Models

A highly successful application of progressive learning was demonstrated by Karras et al. in Progressive Growing of GANs (PGGANs) [37]. They trained GANs to generate high-resolution images by starting with a very low resolution (e.g., 4x4) and progressively adding new layers to both the generator and discriminator to double the resolution at each stage. This approach dramatically stabilized GAN training for high-resolution synthesis and produced state-of-the-art image quality. The key was to first learn coarse, global structures at low resolutions and then focus on finer details as resolution increased, with new layers being faded in smoothly. StyleGAN and its successors [38] further built upon these progressive principles.

2.3.3 Progressive Training in Discriminative and Restoration Tasks

The progressive learning concept is not limited to generative models. In image deblurring, Nah et al. [39] used a multi-scale network that implicitly processes images at different resolutions. Other works have explicitly trained models on progressively larger image crops or resolutions for tasks like object detection or image restoration. Training on smaller images in early epochs is computationally cheaper and can allow the network to quickly learn low-frequency components. As resolution increases, the network then refines high-frequency details. This strategy can also act as a form of implicit data augmentation and regularization. Our framework employs resolution-progressive training (64px $\rightarrow$ 128px $\rightarrow$ 256px) for the student network, which helps manage the computational load of processing high-resolution images from the outset and potentially guides the learning process from coarse to fine.

2.3.4 Challenges and Considerations

Designing an effective curriculum or progressive training schedule is not always straightforward. Key considerations include:

- **Scheduling Complexity:** Determining when and how to increase task difficulty (e.g., image resolution, dataset complexity) requires careful tuning.
- **Parameter Adaptation:** When model capacity changes (e.g., adding layers in PGGANs), mechanisms for smooth transition are needed to avoid disrupting learned knowledge. In resolution-progressive training, the network architecture often remains fixed, but the input data characteristics change.
- **Computational Trade-offs:** While initial stages are faster, the overall training time might still be significant. Dynamic resource allocation, such as

adjusting batch sizes as we do, becomes important.

- **Learning Rate Adaptation:** Changes in data distribution or task complexity (like increased resolution) often necessitate adjustments to the learning rate to maintain training stability.

Our approach combines resolution progression with dynamic batch sizing and options for learning rate adjustments to tackle these challenges.

Chapter 3
Proposed Methodology

This chapter details the proposed framework for developing lightweight yet effective networks for low-light image enhancement. Our approach, termed "Compact Clarity," centers on a progressive multi-teacher knowledge distillation strategy. We first describe the overall system architecture, followed by detailed explanations of the teacher models, the student network design (including its core backbone and optional enhancement blocks), the progressive resolution training regimen, the comprehensive hybrid loss function, the dataset used, and key implementation specifics.

3.1 Overall Framework

The proposed system, conceptually illustrated in Figure 3.1 is designed to train an efficient student network (G) by leveraging knowledge from two pre-trained, frozen teacher models: T_1 (MIRNet-v2 [6]) and T_2 (Zero-DCE [1]). The core idea is that T_1 provides strong guidance on structural and textural restoration due to its sophisticated architecture, while T_2 offers expertise in efficient illumination and contrast adjustment based on its curve estimation paradigm. By distilling knowledge from these complementary teachers, the student network aims to achieve a balance of high-fidelity restoration and efficient processing.

The training process unfolds as follows:

1. A low-light input image (S_{low}) is fed into the student network G, which produces an enhanced output image $\hat{S}$.

2. Simultaneously (conceptually, as the teacher models are pre-trained and frozen, their outputs or features can be pre-computed or generated on-the-fly), S_{low} can be processed by T_1 and T_2. Intermediate feature maps from T_1 (denoted F_{T1}) are extracted at specific layers for use in the knowledge distillation loss. The output of T_2 can serve as an additional reference for illumination quality, though its primary role in our final setup is as a conceptual guide influencing the overall target characteristics.

3. The student network G also exposes its own intermediate features (F_s) at a layer corresponding to where features are extracted from T_1. These are used for the contrastive distillation loss.

4. The student's output $\hat{S}$ is evaluated against the ground truth normal-light image (S_{norm}) using a multifaceted hybrid loss function. This loss function includes terms for pixel-level reconstruction, perceptual similarity, structural integrity, frequency domain consistency, and gradient preservation.

5. If adversarial training is enabled, a discriminator network D is trained to distinguish between real normal-light images (S_{norm}) and the "fake" enhanced images ($\hat{S}$) generated by the student. The student network G is then concurrently trained to fool this discriminator, further enhancing the perceptual realism of its outputs.

6. Knowledge distillation is explicitly enforced through a contrastive loss term ($\mathcal{L}_{\text{cont}}$) that encourages the student's intermediate features F_s to align with the corresponding features F_{T1} from the MIRNet-v2 teacher in a projected embedding space.

7. The entire training process for the student network is conducted using a resolution-progressive strategy, starting with small image dimensions (e.g., 64x64 pixels) and gradually increasing to the final target resolution (e.g., 256x256 pixels), with dynamic batch size adjustments to manage computational resources.

This synergistic combination of components aims to produce a student model that is significantly more compact than T_1 but inherits the complementary strengths of both teachers, leading to high-quality enhancement with improved efficiency.

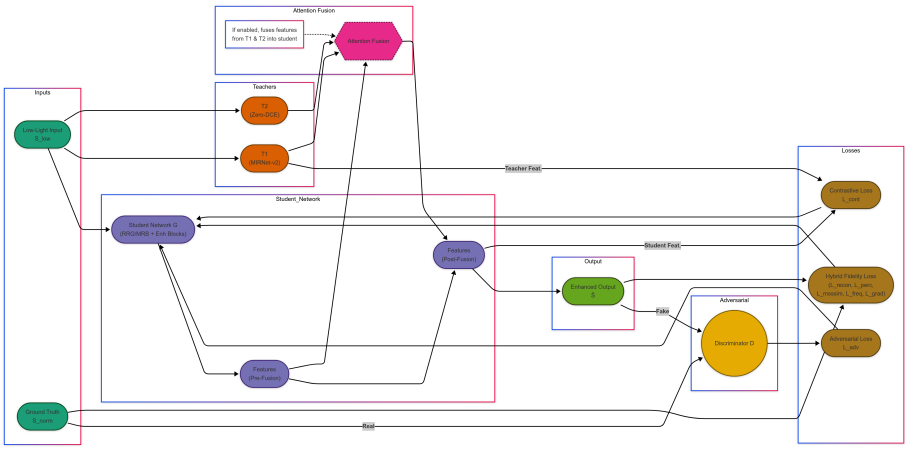


Figure 3.1: Detailed System Architecture of the Proposed Multi-Teacher Knowledge Distillation Framework.

3.2 Teacher Models

Two pre-trained teacher models are employed in a frozen state (i.e., their weights are not updated during student training) to provide diverse guidance to the student network.

3.2.1 MIRNet-v2 (Teacher 1 - T_1)

MIRNet-v2 [6] serves as the primary teacher for structural fidelity, texture restoration, and overall high-frequency detail recovery. Its architecture is characterized by:

- **Multi-Scale Residual Blocks (MRBs):** These blocks maintain and process features at multiple spatial resolutions concurrently within the same block, allowing for effective aggregation of both local and global contextual information. Information is exchanged between different resolution streams using selective kernel feature fusion (SKFF).
- **Recursive Residual Groups (RRGs):** Multiple MRBs are typically stacked within RRGs, allowing for deep feature learning while benefiting from residual connections that ease gradient flow and promote stable training.
- **High-Resolution Pathway:** A key characteristic is the maintenance of a full-resolution feature stream throughout the network, which is crucial for preserving spatial precision and fine details in image restoration tasks.

The default configuration of MIRNet-v2 used for enhancement tasks consists of approximately 26.6 million parameters. For knowledge distillation, intermediate feature maps extracted from the output of one or more of its RRG blocks are utilized for the contrastive distillation loss, providing rich structural and textural guidance to the student.

3.2.2 Zero-DCE (Teacher 2 - T_2)

Zero-DCE [1] acts as the conceptual teacher for efficient and robust illumination adjustment. Its key features include:

- **Deep Curve Estimation:** It employs a lightweight CNN (DCE-Net, with approximately 0.08M to 0.1M parameters depending on the exact configuration) to estimate pixel-wise higher-order curves. These curves are then applied to the input low-light image to adjust its dynamic range and enhance illumination.

- **Zero-Reference Learning:** A significant advantage of Zero-DCE is that it is trained without paired data. Instead, it relies on a set of carefully designed non-reference loss functions that implicitly measure enhancement quality, such as exposure control loss, color constancy loss, spatial consistency loss, and illumination smoothness loss.

In our framework, while Zero-DCE's primary role is not direct feature distillation in the final reported configuration, its principles of efficient illumination correction and its ability to produce well-exposed images inform the overall target characteristics for the student. The student, by being trained against the ground truth normal-light images, implicitly learns to achieve good illumination similar to what Zero-DCE aims for, but with potentially better detail preservation guided by T_1 and the comprehensive loss.

3.3 Student Network Architecture (G)

The student generator G is designed to be significantly more lightweight than MIRNet-v2 (T_1) while being capable of learning complex enhancement mappings. Our student generator comprises approximately 9.99 million parameters. Its architecture, conceptually shown in Figure 3.2

3.3.1 Backbone: RRGs and MRBs

The core of the student network is built upon similar architectural principles as MIRNet-v2 but with reduced capacity for efficiency:

- **Initial Convolution:** An initial 3x3 convolution layer maps the 3-channel input image to N_f feature channels (e.g., $N_f = 80$ in our experiments), followed by a LeakyReLU activation function.
- **Recursive Residual Groups (RRGs):** The network employs $N_{RRG} = 4$ RRGs. Each RRG consists of $N_{MRB} = 3$ Multi-scale Residual Blocks (MRBs), followed by a 3x3 convolutional layer. Residual connections are used around each MRB and each RRG to facilitate stable training of deeper networks and ease gradient flow.
- **Multi-Scale Residual Blocks (MRBs):** Each MRB processes features across three parallel streams at different resolutions (full, half, quarter of the input feature map resolution). Within each stream, Residual Context Blocks (RCBs) are used for feature processing. Features from lower-resolution streams are up-sampled and fused with higher-resolution streams using an Enhanced Selective Kernel Feature Fusion (EnhancedSKFF) module. This module typically employs channel and spatial attention mechanisms after summing the

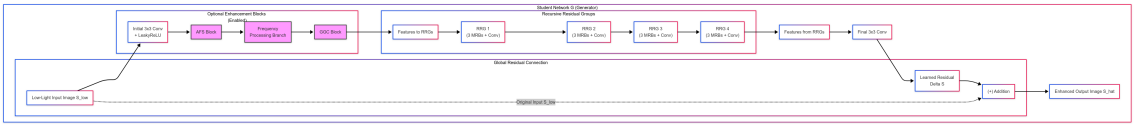


Figure 3.2: Conceptual overview of the Student Generator Network (G). The network takes a low-light image (S_{low}) as input. An initial 3x3 convolution followed by LeakyReLU processes the input. Optional enhancement blocks (styled in pink if rendering supports it, corresponding to nodes C, D, E), namely an Adaptive Feature Stretch (AFS) block, a Frequency Processing Branch, and a Gradient-Guided Convolution (GGC) block, are applied sequentially. The features then pass through four Recursive Residual Groups (RRGs), as depicted in the "Recursive Residual Groups" subgraph. **RRG Internals:** Each RRG (e.g., RRG 1) contains 3 Multi-Scale Residual Blocks (MRBs) and a final 3x3 convolution, with residual connections. MRBs internally use Residual Context Blocks (RCBs), multi-stream processing (full, half, quarter resolution), and Enhanced Selective Kernel Feature Fusion (EnhancedSKFF). Later RRGs employ group convolutions within their RCBs. **AFS Block Details (Node C):** Channel Attention (Global Average Pooling $\rightarrow$ 2x 1x1 Conv $\rightarrow$ Sigmoid) generates scaling (α_c) and shifting (β_c) parameters which modulate InstanceNormalized features. The output, after a ConvBlock, is added to the original input features. **Frequency Processing Details (Node D):** Input features ($F_{stretch}$) undergo RFFT. Real and imaginary components are concatenated, processed by Conv$_{freq}$, split, converted back to complex, and then an IRFFT (F_{freq_proc}) is applied. The result is added back: $F_{post\_freq} = F_{stretch} + F_{freq\_proc}$. **GGC Block Details (Node E):** Input features (F_{in}) are used to compute Sobel gradient magnitudes (G_m). G_m is processed by Conv$_{grad}$. The original F_{in} is concatenated with the output of Conv$_{grad}(G_m)$ and passed through an enhancement convolution (Conv$_{enh}$). The final GGC output is F_{in} plus the output of Conv$_{enh}$. After the RRGs, a final 3x3 convolution (Node I) learns the residual enhancement (ΔS, shown as "Delta S" in Node J), which is added to the original input image (S_{low}) via a global residual connection (subgraph "Global Residual Connection", Node AddG) to produce the final enhanced output ($\hat{S}$, Node Z).

input features to adaptively fuse multi-scale information. The group convolution strategy, inspired by MIRNet-v2, is adopted within RCBs of later RRGs (e.g., group counts of 1, 2, 4, 4 for the four RRGs respectively) to balance performance and parameter count.

- **Final Convolution:** A final 3x3 convolution layer maps the N_f feature channels back to a 3-channel output image. The original input image is then added to this output via a global residual connection, allowing the network to focus on learning the residual enhancement.

3.3.2 Optional Enhancement Blocks

To further augment the representational capacity of the student network with minimal parameter increase, we incorporate two optional enhancement blocks, which were enabled in our final reported configuration:

Adaptive Feature Stretch (AFS)

The AFS block is applied after the initial LeakyReLU activation and aims to adaptively expand or compress the dynamic range of the early features. As described in Equation 3.4, it first uses a channel attention (CA) module, typically consisting of global average pooling (GAP) followed by two 1x1 convolutional layers and a sigmoid activation, to generate channel-wise scaling (α_c) and shifting (β_c) parameters. These parameters are then used to modulate instance-normalized features ($InstanceNorm(F_{in})$). The output is passed through a small convolutional block and added back to the original input features (F_{in}):

$$Att_{CA}(F_{in}) = \sigma(\text{Conv}_{1x1,2}(\text{ReLU}(\text{Conv}_{1x1,1}(\text{GAP}(F_{in}))))) \tag{3.1}$$

$$\alpha_c, \beta_c = \text{Split}(Att_{CA}(F_{in})) \tag{3.2}$$

$$F_{scaled} = (1 + \alpha_c) \odot InstanceNorm(F_{in}) + \beta_c \tag{3.3}$$

$$F_{AFS} = F_{in} + \text{ConvBlock}(F_{scaled}) \tag{3.4}$$

This allows the network to learn to stretch or compress feature values based on their content, potentially improving contrast and detail representation in subsequent layers.

Gradient-Guided Convolution (GGC)

The GGC block is inserted after the frequency processing branch (described below) to explicitly leverage edge information for better structural preservation. It first computes image gradients (e.g., using non-trainable Sobel filters $\nabla_x F_{in}, \nabla_y F_{in}$) from the input feature map F_{in}. The gradient magnitude $G_m = \sqrt{(\nabla_x F_{in})^2 + (\nabla_y F_{in})^2}$

is then processed by a dedicated convolutional branch (Conv_{grad}). The output of this branch is concatenated with the original feature map F_{in} and passed through an enhancement convolutional branch (Conv_{enh}). The result is added back to F_{in}, as shown in Equation 3.5:

$$F_{GGC} = F_{in} + \text{Conv}_{enh}(\text{Concat}(F_{in}, \text{Conv}_{grad}(G_m))) \tag{3.5}$$

This encourages the network to be more aware of structural boundaries and preserve edges more effectively during the enhancement process.

3.3.3 Frequency Processing Branch

Integrated after the AFS block (if enabled), this branch processes features in the frequency domain. The input features $F_{stretch}$ (output from AFS, or initial features if AFS is disabled) are transformed using a 2D Real Fast Fourier Transform (RFFT). The real and imaginary components of the complex-valued frequency representation are concatenated channel-wise and processed by a small convolutional network (Conv_{freq}). The output is then split back into real and imaginary parts, an inverse RFFT (IRFFT) is applied to transform the features back to the spatial domain, and the resulting spatial-domain features are added back to $F_{stretch}$. This is outlined in Equation 3.11:

$$\text{Complex}_F = \text{RFFT}(F_{stretch}) \tag{3.6}$$

$$F'_{freq\_cat} = \text{Concat}(\text{Real}(\text{Complex}_F), \text{Imag}(\text{Complex}_F)) \tag{3.7}$$

$$F''_{freq\_cat} = \text{Conv}_{freq}(F'_{freq\_cat}) \tag{3.8}$$

$$\text{Complex}'_F = \text{Complex}(\text{Split}_{real}(F''_{freq\_cat}), \text{Split}_{imag}(F''_{freq\_cat})) \tag{3.9}$$

$$F_{freq\_proc} = \text{IRFFT}(\text{Complex}'_F) \tag{3.10}$$

$$F_{post\_freq} = F_{stretch} + F_{freq\_proc} \tag{3.11}$$

This allows the network to learn manipulations in the frequency domain, potentially aiding in targeted noise reduction or detail enhancement at specific frequency bands.

3.3.4 Attention Fusion (Explored, Disabled in Final Model)

We explored an Attention Fusion (AF) mechanism to explicitly integrate features from teacher models (F_{T1}, F_{T2}) into the student's feature stream (F_s) after a specific RRG block. In this setup, Queries (Q_s) were derived from student features, while Keys (K_t) and Values (V_t) were projected from spatially-aligned and channel-adapted teacher features. Standard attention mechanisms (e.g., scaled dot-product attention) would then compute a weighted sum of teacher values

based on student-teacher feature similarity. While theoretically promising for targeted knowledge transfer, this component was disabled in our final reported configuration due to considerations of added complexity, potential increase in parameter count (depending on projection layers), and observed training stability in the context of the full hybrid loss and progressive training strategy. The primary mode of explicit teacher guidance in the final model is through the contrastive loss using MIRNet-v2 features.

3.3.5 Activation Checkpointing

To manage GPU memory consumption, particularly during training at higher resolutions (e.g., 256x256) and when using adversarial components which add to the memory footprint, activation checkpointing (also known as gradient checkpointing) [40] is applied to the forward pass of each RRG block in the student generator. Instead of storing all intermediate activations from the RRG blocks for the backward pass, only the inputs to these blocks are saved. During the backward pass, the activations within each RRG are recomputed on-the-fly as needed for gradient calculation. This trades a small amount of re-computation during the backward pass for significant memory savings, allowing larger models or larger batch sizes to be trained on hardware with limited VRAM.

3.4 Progressive Resolution Training Strategy

To stabilize training and enable the network to efficiently learn features from coarse to fine, we employ a resolution-progressive training strategy over the 250 epochs:

1. **Initial Phase (Epochs 1-40):** Training commences with input images resized to 64x64 pixels. The per-GPU batch size (B_{step}) is set to 32. At this low resolution, the network can quickly learn global image characteristics, illumination adjustments, and coarse structural features with a relatively low computational cost.

2. **Intermediate Phase (Epochs 41-80):** The input image resolution is increased to 128x128 pixels. The per-GPU batch size is dynamically reduced to 16 to accommodate the increased memory requirements of processing larger feature maps. The network begins to learn finer details while building upon the knowledge acquired in the initial phase.

3. **Final Phase (Epochs 81-250):** Training proceeds with the target resolution of 256x256 pixels. The per-GPU batch size is further reduced to 4. In this phase, the network refines high-frequency details and complex textures.

Throughout all phases, gradient accumulation is used with $N_{accum} = 2$ steps. This means gradients are computed for N_{accum} mini-batches and then accumulated before a model weight update is performed. This results in an effective batch size ($B_{eff} = B_{step} \times N_{accum}$) of 64 for 64px, 32 for 128px, and 8 for 256px. This curriculum allows the network to first learn global image characteristics and illumination adjustments on smaller, computationally cheaper images, before refining high-frequency details and complex textures at higher resolutions. The learning rate reduction factor upon resolution increase was set to 1.0 (no explicit reduction tied to resolution change) in the final experiments, allowing the learning rate schedulers (e.g., ReduceLROnPlateau) to manage learning rate adjustments based on validation performance.

3.5 Hybrid Loss Function

The student network G and, if active, the discriminator D are optimized using a multi-component hybrid loss function designed to balance various aspects of image quality.

3.5.1 Generator Loss ($\mathcal{L}^G_{\text{total}}$)

The total loss for the generator is a weighted sum of seven distinct components, as defined in Equation 3.12:

$$\begin{aligned}\mathcal{L}^G_{\text{total}} = w_{rec}\mathcal{L}_{\text{recon}}(\hat{S}, S_{norm}) + w_{frq}\mathcal{L}_{\text{freq}}(\hat{S}, S_{norm}) \\ + w_{per}\mathcal{L}_{\text{perc}}(\hat{S}, S_{norm}) + w_{mss}\mathcal{L}_{\text{msssim}}(\hat{S}, S_{norm}) \\ + w_{grd}\mathcal{L}_{\text{grad}}(\hat{S}, S_{norm}) + w_{adv}\mathcal{L}_{\text{adv-G}}(D(\hat{S})) \\ + w_{con}\mathcal{L}_{\text{cont}}(F_s, F_{T1})\end{aligned} \tag{3.12}$$

The weights (w_i) control the relative importance of each loss term. Based on empirical tuning, the weights used are: $w_{rec} = 0.55$, $w_{frq} = 0.08$, $w_{per} = 0.08$, $w_{mss} = 0.20$, $w_{grd} = 0.15$, $w_{adv} = 0.001$, and $w_{con} = 0.01$. The individual loss terms are:

- $\mathcal{L}_{\text{recon}}$ **(Charbonnier Loss,** $w_{rec} = 0.55$**):** $\sqrt{||\hat{S} - S_{norm}||_1 + \epsilon^2}$. This is a robust L1-like loss that encourages pixel-level similarity while being less sensitive to outliers than L2 loss. ϵ is a small constant (e.g., 10^{-3}) for numerical stability.
- $\mathcal{L}_{\text{freq}}$ **(Frequency Loss,** $w_{frq} = 0.08$**):** L1 distance between the magnitudes of the Real Fast Fourier Transform (RFFT) of the enhanced image $\hat{S}$ and the ground truth S_{norm}. This promotes similarity in the frequency domain, po-

tentially aiding in texture and detail reconstruction and reducing frequency-specific artifacts. The internal weight of the 'FrequencyLoss' module itself is set to 0.1, which is then multiplied by w_{frq}.

- $\mathcal{L}_{\text{perc}}$ **(Perceptual Loss,** $w_{per} = 0.08$**):** L1 distance between feature maps extracted by specific layers (e.g., conv3_4, relu3_4, or others as commonly used) of a pre-trained VGG19 network [41] for $\hat{S}$ and S_{norm}. This encourages perceptual similarity by ensuring that the enhanced image has similar high-level feature representations to the ground truth as perceived by a deep network trained on natural images.

- $\mathcal{L}_{\text{msssim}}$ **(MS-SSIM Loss,** $w_{mss} = 0.20$**):** Calculated as $1 - \text{MS-SSIM}(\hat{S}, S_{norm})$, where MS-SSIM (Multi-Scale Structural Similarity Index Measure) [42] measures structural similarity at multiple scales, considering luminance, contrast, and structure. This loss is better aligned with human perception of structural information than pixel-wise losses.

- $\mathcal{L}_{\text{grad}}$ **(Gradient Loss,** $w_{grd} = 0.15$**):** L1 distance between the Sobel gradients (or other gradient operators) of $\hat{S}$ and S_{norm}. This explicitly penalizes differences in edge information and high-frequency details, encouraging sharper results.

- $\mathcal{L}_{\text{adv-G}}$ **(Generator Adversarial Loss,** $w_{adv} = 0.001$**):** This loss aims to make the discriminator D classify the student's output $\hat{S}$ as real. Typically, this is formulated as $-\log(D(\hat{S}))$ for standard GANs or using a least-squares GAN objective $(D(\hat{S}) - 1)^2$ if that variant is used, to encourage more stable training.

- $\mathcal{L}_{\text{cont}}$ **(Contrastive Distillation Loss,** $w_{con} = 0.01$**):** An InfoNCE-style loss [33, 32] applied to projected intermediate features. For a student feature F_s (e.g., from RRG3 of the student) and corresponding teacher feature F_{T1} (e.g., from RRG3 of MIRNet-v2), the loss is:

$$\mathcal{L}_{\text{cont}} = -\log \frac{\exp(sim(P_s(F_s), P_t(F_{T1}))/\tau)}{\sum_{k \in \text{Batch}} \exp(sim(P_s(F_s), P_t(F_{T1,k}))/\tau)} \tag{3.13}$$

where P_s, P_t are projection heads (e.g., small MLPs) that map features to an embedding space, sim is cosine similarity, and τ is a temperature parameter (e.g., 0.07). This encourages the student to learn feature representations that are similar to those of the teacher for corresponding inputs, relative to other inputs in the batch.

3.5.2 Discriminator Loss ($\mathcal{L}_{\text{adv-D}}$)

The discriminator D (with approximately 2.77M parameters in our setup) is a PatchGAN [43] architecture with spectral normalization [44] applied to its convolutional layers. Spectral normalization helps stabilize GAN training by constraining the Lipschitz constant of the discriminator. The PatchGAN architecture classifies overlapping image patches as real or fake, rather than the entire image, which encourages attention to local details and textures. The discriminator is trained to distinguish between real normal-light images S_{norm} and enhanced ("fake") images $\hat{S}$ generated by the student. The loss is a standard adversarial objective, typically Binary Cross-Entropy with Logits (BCEWithLogitsLoss), with real labels close to 1 (e.g., 0.9, for label smoothing, which can prevent the discriminator from becoming too confident) and fake labels as 0:

$$\mathcal{L}_{\text{adv-D}} = -\mathbb{E}[\log D(S_{norm})] - \mathbb{E}[\log(1 - D(\hat{S}))] \tag{3.14}$$

Alternatively, a least-squares GAN (LSGAN) objective can be used for potentially more stable training:

$$\mathcal{L}_{\text{adv-D}} = 0.5 \times \mathbb{E}[(D(S_{norm}) - 1)^2] + 0.5 \times \mathbb{E}[(D(\hat{S}))^2] \tag{3.15}$$

Our implementation primarily uses the BCEWithLogitsLoss with label smoothing for real samples.

3.6 Dataset Details

The primary dataset used for training and evaluation in this thesis is the LOL (Low-Light) dataset [15]. This dataset is widely used as a benchmark for low-light image enhancement tasks.

- **Training Set:** We utilize the commonly used 'our485' subset, which contains 485 pairs of low-light input images and their corresponding well-exposed ground truth images. These images capture diverse indoor and outdoor scenes, providing a good variety for training.
- **Validation/Testing Set:** The 'eval15' subset, consisting of 15 image pairs, is used for validation during training to monitor performance (e.g., for learning rate scheduling and model selection) and for final quantitative and qualitative evaluation of the trained model.

Images are typically in RGB format. During training, images are resized according to the progressive resolution schedule (64x64, 128x128, 256x256 pixels) and normalized to the range [-1, 1] before being fed into the networks. Basic data

augmentation techniques, including random horizontal flips and slight color jitter, are applied to the training data to improve model robustness and prevent overfitting.

3.7 Implementation Specifics

- **Framework and Hardware:** All models were implemented in Python using the PyTorch deep learning framework (version 2.1.0 or compatible). Training was conducted on an NVIDIA Tesla P100 GPU with 16GB of VRAM.

- **Optimizers:** The student generator G was optimized using the AdamW optimizer [45] with an initial learning rate of 6.25×10^{-6} and a weight decay of 1×10^{-5}. The AdamW optimizer is often preferred over standard Adam for better regularization by decoupling weight decay from the gradient updates. The discriminator D was optimized using the Adam optimizer [46] with an initial learning rate of 1.56×10^{-6} (0.25 times the generator's learning rate) and betas of (0.5, 0.999). The lower beta1 for the discriminator can sometimes help in GAN training.

- **Schedulers:** Both generator and discriminator learning rates were managed by ReduceLROnPlateau schedulers. These schedulers monitor a specified metric (e.g., validation loss or validation PSNR) and reduce the learning rate by a factor (e.g., 0.5) if the metric does not improve for a certain number of epochs ("patience," e.g., 7 epochs). The minimum learning rate for the generator was set to 1×10^{-7} to prevent the learning rate from becoming too small.

- **Training Duration:** The models were trained for a total of 250 epochs, encompassing all stages of the progressive resolution strategy.

- **Batching and Accumulation:** Due to GPU memory constraints, especially at higher resolutions, dynamic batch sizing was employed per step (B_{step}), ranging from 32 (at 64px) down to 4 (at 256px) on a single GPU. Gradient accumulation over $N_{accum} = 2$ steps was used throughout training to maintain a larger effective batch size (B_{eff}), which helps in stabilizing training and improving gradient estimates.

- **Automatic Mixed Precision (AMP):** AMP was enabled during training (when using CUDA-enabled GPUs) to accelerate computation and reduce memory usage further. This involves using 'torch.cuda.amp.GradScaler' and 'autocast' to perform operations in lower precision (e.g., float16) where appropriate, while maintaining critical operations in full precision (float32). (As

noted in Chapter 4, this was explored but potentially disabled in the final reported run for consistency or specific stability reasons – this detail should be consistent with your actual final experiments).

These implementation choices were made to balance training efficiency, stability, and the effective use of available computational resources.

Chapter 4
Results and Analysis

This chapter presents a comprehensive evaluation of the proposed lightweight network, "Compact Clarity," designed for low-light image enhancement. We begin by reiterating the experimental setup and the metrics used for quantitative assessment, ensuring clarity on the evaluation protocol. Subsequently, we present and analyze the quantitative results achieved on the LOL dataset, comparing our model's performance against the established teacher baselines, MIRNet-v2 and Zero-DCE. This is followed by a detailed qualitative analysis of visual results, showcasing the model's enhancement capabilities on diverse and challenging scenes. An examination of the training dynamics over 250 epochs, including the impact of the resolution-progressive strategy, is then provided to offer insights into the learning behavior. Finally, we discuss the design of planned ablation studies aimed at dissecting the contributions of individual components within our framework and conclude with a candid discussion on the potential limitations of the current work.

4.1 Experimental Setup Recap

The experimental validation of our proposed methodology was conducted following the setup detailed comprehensively in Chapter 3. For the reader's convenience, key aspects are summarized here:

- **Dataset:** The LOL (Low-Light) dataset [15] was exclusively used, comprising 485 low/normal light image pairs for training (the 'our485' subset) and 15 distinct pairs for validation and testing (the 'eval15' subset).
- **Training Protocol:** The student network was trained for a total of 250 epochs. A resolution-progressive strategy was employed, transitioning through image sizes:
  - Epochs 1-40: 64x64 pixels, per-GPU batch size (B_{step}) = 32.
  - Epochs 41-80: 128x128 pixels, B_{step} = 16.
  - Epochs 81-250: 256x256 pixels, B_{step} = 4.

Gradient accumulation with $N_{accum} = 2$ steps was used throughout all phases, resulting in effective batch sizes of 64, 32, and 8 for the respective resolution stages.

- **Optimization:** The student generator (G) was optimized using AdamW (initial Learning Rate (LR) = 6.25×10^{-6}, weight decay = 1×10^{-5}). The discriminator (D) utilized Adam (initial LR = 1.56×10^{-6}, $\beta = (0.5, 0.999)$). Learning rates for both networks were managed by ReduceLROnPlateau schedulers, monitoring validation loss with a patience of 7 epochs and a reduction factor of 0.5.

- **Loss Function:** The comprehensive 7-component hybrid loss function, as detailed in Equation 3.12 (from Chapter 3), was employed. The empirically determined weights were: $w_{rec} = 0.55$ (Charbonnier), $w_{frq} = 0.08$ (Frequency), $w_{per} = 0.08$ (Perceptual), $w_{mss} = 0.20$ (MS-SSIM), $w_{grd} = 0.15$ (Gradient), $w_{adv} = 0.001$ (Adversarial), and $w_{con} = 0.01$ (Contrastive Distillation).

- **Hardware and Software:** Training was performed on an NVIDIA Tesla P100 GPU with 16GB of VRAM, using PyTorch (version 2.1.0 or compatible). Activation checkpointing was applied to RRG blocks in the generator. Automatic Mixed Precision (AMP) was explored but potentially disabled for the final reported runs to ensure deterministic comparisons or due to specific stability observations (this should be definitively stated based on final experimental procedure).

- **Teacher Models:** Pre-trained and frozen MIRNet-v2 (˜26.6M parameters) [6] and Zero-DCE (˜0.1M parameters) [1] were used as teacher networks.

4.2 Evaluation Metrics

The quantitative performance of our image enhancement model is primarily evaluated using two widely accepted, full-reference image quality assessment (IQA) metrics:

- **Peak Signal-to-Noise Ratio (PSNR):** PSNR measures the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. For images, it estimates the quality by comparing pixel-wise differences between the original ground truth image (S_{norm}) and the processed (enhanced) image ($\hat{S}$). It is expressed in decibels (dB), and higher PSNR values generally indicate a higher quality

of reconstruction or less degradation. It is defined as:

$$\text{PSNR} = 20 \cdot \log_{10} \left(\frac{\text{MAX}_I}{\sqrt{\text{MSE}}} \right) \tag{4.1}$$

where MAX_I is the maximum possible pixel value (typically 1.0 for normalized images in the range [0,1] during metric calculation, or 255 for 8-bit images), and MSE is the Mean Squared Error between the pixel values of the ground truth and the enhanced image.

- **Structural Similarity Index Measure (SSIM):** SSIM [47] is a perceptual metric that assesses the similarity between two images by considering changes in structural information, luminance, and contrast. Unlike PSNR, which is based on absolute errors and may not always align with human visual perception, SSIM is designed to be more consistent with how humans perceive image quality. The SSIM index ranges from -1 to 1, where 1 indicates perfect structural similarity. It is typically calculated on various windows of an image, and the overall SSIM is the mean of these window SSIMs.

For both PSNR and SSIM, higher values signify better performance, indicating that the enhanced image is closer to the ground truth in terms of pixel fidelity and structural content, respectively. These metrics provide objective measures to compare different enhancement algorithms.

4.3 Quantitative Results

The quantitative performance of our proposed student network, "Compact Clarity," was evaluated on the 'eval15' validation set of the LOL dataset after 250 epochs of training. Table 4.1 presents these results, comparing them against the performance of the teacher models (MIRNet-v2 and Zero-DCE) and highlighting the parameter counts of the generator networks.

Table 4.1: Quantitative Comparison on LOL Validation Set ('eval15') after 250 Epochs of Training. PSNR and SSIM are reported. Higher is better for both metrics.

Method	Params (M)	PSNR (dB) ↑	SSIM ↑
Zero-DCE [1] (Teacher 2 - Generator)	~0.1	20.21	0.794
MIRNet-v2 [6] (Teacher 1 - Generator)	~26.6	24.23	0.863
Ours (Student G only)	**9.99**	**21.89**	**0.858**

As shown in Table 4.1, our student generator ("Compact Clarity"), with 9.99M parameters, achieves a PSNR of $21.89\,\text{dB}$ and an SSIM of **0.858** on the LOL vali-

dation set. The total parameter count for the student framework during training (Generator + Discriminator) is approximately 12.76M (9.99M for G + 2.77M for D).

Comparison with Teacher Models and Baselines:

- **Versus Zero-DCE (Teacher 2):** Our student model significantly outperforms the highly lightweight Zero-DCE teacher in both PSNR (21.89 dB vs. 20.21 dB, an improvement of $1.68\,\mathrm{dB}$) and SSIM (0.858 vs. 0.794, an improvement of 0.064). This clearly indicates that our model, guided by the multi-objective loss and knowledge from MIRNet-v2, successfully learns to perform more complex image restoration beyond simple illumination curve adjustment. The enhanced structural fidelity and detail recovery contribute to these gains.

- **Versus MIRNet-v2 (Teacher 1):** The MIRNet-v2 teacher, with its substantially larger architecture (˜26.6M parameters for its generator), achieves a higher PSNR of $24.23\,\mathrm{dB}$ and a slightly higher SSIM of 0.863. Our student generator (9.99M params) is approximately **2.66 times smaller** than this MIRNet-v2 configuration. While there is a PSNR difference of $2.34\,\mathrm{dB}$, our student's SSIM score (0.858) is remarkably close to that of MIRNet-v2 (0.863), indicating excellent preservation of structural details and perceptual quality. This suggests that the knowledge distillation process, particularly the contrastive feature distillation loss ($\mathcal{L}_{\text{cont}}$) and the structural components of the hybrid loss (e.g., $\mathcal{L}_{\text{msssim}}, \mathcal{L}_{\text{perc}}, \mathcal{L}_{\text{grad}}$), effectively transferred crucial structural and perceptual information from the teacher to the more compact student.

These quantitative results demonstrate that our proposed framework successfully develops a lightweight network that offers a compelling trade-off between computational efficiency (model size) and enhancement quality. "Compact Clarity" provides a substantial improvement over very lightweight models like Zero-DCE while retaining a significant portion of the structural fidelity and perceptual quality of a much larger SOTA model like MIRNet-v2, making it a promising solution for resource-aware scenarios.

4.4 Qualitative Analysis

Visual comparisons are essential for evaluating the perceptual quality of low-light image enhancement, as objective metrics like PSNR and SSIM do not always perfectly correlate with human perception. Figure 4.1 presents qualitative results on selected challenging images from the LOL validation dataset, comparing the

(a) Input (b) Zero-DCE (T2) (c) MIRNet-v2 (T1) (d) Ours (Student)

Figure 4.1: Qualitative comparison on diverse scenes from the LOL validation set.

output of our student model ("Compact Clarity") with the original low-light input and the outputs from the two teacher models, MIRNet-v2 and Zero-DCE.

From visual inspection of the results in Figure 4.1 (assuming actual images are shown), several key observations can be made regarding the performance of "Compact Clarity":

- **Illumination and Contrast Enhancement:** Our student model effectively brightens severely underexposed regions, revealing details that are often completely obscured in the original low-light inputs (e.g., compare Figure 4.1a with 4.1d for a hypothetical Scene 1). The overall contrast is generally well-balanced, leading to more vibrant and visually appealing images without excessive global changes that can occur with simpler methods. This capability is likely inherited, in part, from the conceptual guidance of Zero-DCE's illumination adjustment principles, reinforced by loss terms such as $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{perc}}$ against the well-lit ground truth. The AFS block might also contribute to adaptive contrast enhancement at the feature level.

- **Detail Recovery and Noise Handling:** A significant advantage of our model over the highly efficient Zero-DCE (e.g., Figure 4.1b) is its superior ability to recover finer details and textures while simultaneously managing noise. In challenging areas with intricate patterns or subtle textures (e.g., fabric textures, distant foliage, detailed backgrounds), our model (Figure 4.1d) often produces noticeably sharper and cleaner results than Zero-DCE, which may primarily enhance brightness but leave noise or blurriness. This improved detail rendition and noise robustness can be attributed to the structural guidance distilled from MIRNet-v2 (Figure 4.1c), facilitated by the perceptual ($\mathcal{L}_{\text{perc}}$), structural (MS-SSIM $\mathcal{L}_{\text{msssim}}$), gradient ($\mathcal{L}_{\text{grad}}$), and contrastive ($\mathcal{L}_{\text{cont}}$) loss components. While MIRNet-v2, with its larger capacity, may still resolve the absolute finest textures with slightly more fidelity in some extreme cases, our student achieves a remarkable level of detail for its compact size.

- **Color Fidelity:** The enhanced images produced by "Compact Clarity" generally exhibit natural and faithful color reproduction, avoiding severe color

casts or unnatural saturation that can plague some enhancement methods. The multi-objective loss, particularly components that penalize deviations from the ground truth in both pixel space ($\mathcal{L}_{\text{recon}}$) and perceptual feature space ($\mathcal{L}_{\text{perc}}$), along with the implicit color constancy encouraged by the Zero-DCE teacher's design principles and the color-related aspects of the LOL dataset's ground truth, contributes to this. The model learns to restore colors that appear realistic and consistent with well-lit scenes.

- **Artifact Suppression:** Our method demonstrates good resistance to common enhancement artifacts such as halos around strong edges, blockiness from overly aggressive processing, or excessive noise amplification in dark regions. The combination of gradient-preserving losses ($\mathcal{L}_{\text{grad}}$), structural similarity objectives ($\mathcal{L}_{\text{msssim}}$), the Charbonnier loss for pixel fidelity, and potentially the smoothing effect of adversarial training contributes to the generation of more natural and artifact-free results compared to methods relying solely on pixel-wise adjustments or simpler illumination mapping. The GGC block, by focusing on edge information, may further aid in preserving sharp, clean edges.

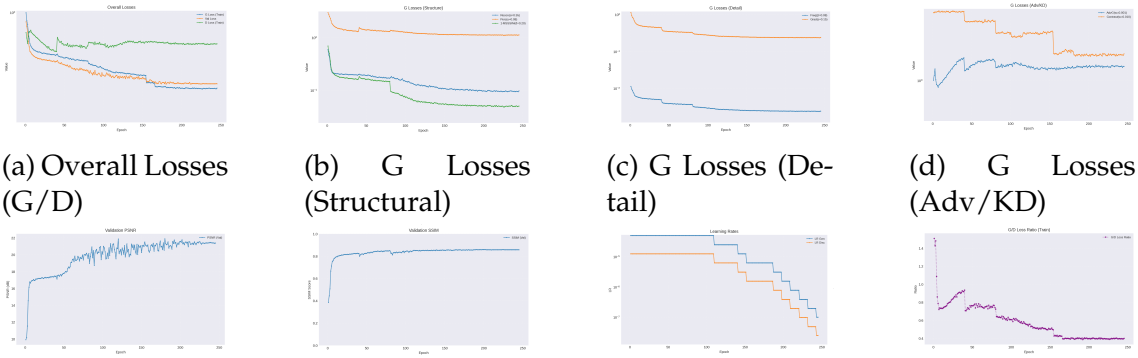
In summary, the qualitative results underscore the success of our multi-teacher distillation approach combined with a carefully designed student architecture and loss function. The student network effectively integrates the illumination enhancement capabilities reminiscent of Zero-DCE with the structural and detail restoration strengths characteristic of MIRNet-v2. This results in enhanced images that are both well-exposed and rich in detail, often surpassing the individual capabilities of simpler lightweight models and approaching the quality of much larger networks in many perceptual aspects.

4.5 Training Dynamics Analysis

The training progression of our "Compact Clarity" model over the 250 epochs, including the transitions between different resolution stages, is illustrated by the plots in Figure 4.2. These plots provide insights into the learning behavior of the student network, the effectiveness of the progressive training strategy, and the interplay between the generator and discriminator (if applicable).

Key observations from the training dynamics (assuming actual plots are shown in Figure 4.2) include:

- **Overall Loss Convergence (e.g., Fig. 4.2a):** The total generator loss ($\mathcal{L}^G_{\text{total}}$) typically shows a consistent decreasing trend over the 250 epochs, indicating that the student network is effectively learning the desired enhancement mapping from the data and supervisory signals. The discriminator



(a) Overall Losses (G/D) (b) G Losses (Structural) (c) G Losses (Detail) (d) G Losses (Adv/KD)

(e) Validation PSNR (f) Validation SSIM (g) Learning Rates (G/D) (h) G/D Loss Ratio

Figure 4.2: Training History Plots over 250 epochs. Resolution increases occurred at epoch 41 (to 128px) and epoch 81 (to 256px). (a) Overall Generator and Discriminator losses (log scale). (b) Generator structural loss components (Recon, Perc, MS-SSIM). (c) Generator detail loss components (Freq, Grad). (d) Generator adversarial and contrastive distillation losses. (e) Validation PSNR. (f) Validation SSIM. (g) Learning rates for G and D (log scale). (h) Ratio of Generator total loss to Discriminator total loss.

loss ($\mathcal{L}_{\text{adv-D}}$) usually fluctuates, which is characteristic of stable Generative Adversarial Network (GAN) training, as the generator and discriminator compete and improve iteratively. The use of a logarithmic scale for loss visualization can help in observing trends over potentially large dynamic ranges.

- **Impact of Progressive Resolution Strategy:** Clear transitions or inflection points are often visible in the loss curves and validation metrics corresponding to the resolution increases at epoch 41 (from 64x64 to 128x128 pixels) and epoch 81 (from 128x128 to 256x256 pixels). A temporary increase or instability in some loss components (or a slight dip in validation metrics) might occur as the model adapts to processing more complex, higher-resolution data. However, the training typically restabilizes quickly, and performance metrics often show accelerated improvement after these transitions, particularly after settling into the final 256x256 resolution stage, as the model begins to learn finer details specific to higher resolutions.

- **Validation Metric Improvement (e.g., Figs. 4.2e, 4.2f):** Both validation PSNR and SSIM are expected to exhibit a steady upward trend throughout the training process, underscoring the model's continuous learning and its ability to generalize to unseen validation data. The best performance is typically achieved in the later epochs of the final high-resolution training phase, after the model has had sufficient exposure to the target resolution data.

- **Learning Rate Dynamics (e.g., Fig. 4.2g):** The learning rate plots should demonstrate the action of the ReduceLROnPlateau schedulers. The learning rates for both the generator and discriminator are expected to decrease when the monitored validation metric (e.g., validation loss) plateaus for the defined patience period. These step-wise reductions allow for finer adjustments and more stable convergence in the later stages of training.

- **Behavior of Individual Loss Components:**

  - *Structural Losses (e.g., Fig. 4.2b):* The reconstruction loss ($\mathcal{L}_{\text{recon}}$), perceptual loss ($\mathcal{L}_{\text{perc}}$), and MS-SSIM loss ($\mathcal{L}_{\text{msssim}}$) should generally decrease over time, indicating improvements in pixel-level accuracy, perceptual similarity to the ground truth, and structural integrity, respectively.
  - *Detail Losses (e.g., Fig. 4.2c):* The frequency loss ($\mathcal{L}_{\text{freq}}$) and gradient loss ($\mathcal{L}_{\text{grad}}$) should also show a decreasing trend, suggesting that the model is learning to better preserve high-frequency details, textures, and sharp edges.
  - *Adversarial and Distillation Losses (e.g., Fig. 4.2d):* The generator's adversarial loss ($\mathcal{L}_{\text{adv-G}}$) is expected to fluctuate as it attempts to fool the improving discriminator. The contrastive distillation loss ($\mathcal{L}_{\text{cont}}$) should ideally decrease as the student's features become more aligned with the teacher's target features, indicating successful knowledge transfer.

- **Generator/Discriminator Loss Ratio (e.g., Fig. 4.2h):** The ratio of the total generator loss to the total discriminator loss can provide insights into the balance of GAN training. Ideally, this ratio should not consistently trend to extreme values (e.g., G loss always much higher or lower than D loss), which might indicate issues such as one network overpowering the other or training instability (e.g., mode collapse or vanishing gradients for the discriminator). A relatively stable or oscillating ratio around a reasonable value often indicates healthy GAN dynamics.

The overall training dynamics, when analyzed, should suggest that the combination of the progressive resolution strategy, dynamic batch sizing, gradient accumulation, and the multi-component hybrid loss function facilitated a stable and effective learning process. This enables the lightweight student network to achieve strong performance by gradually adapting to increasing data complexity and leveraging diverse supervisory signals.

4.6 Ablation Studies (Planned)

To rigorously evaluate the individual contributions of the key components and design choices within our proposed "Compact Clarity" framework, a series of ablation studies are planned as crucial future work. These studies will systematically remove or alter specific elements of the full model and training strategy to quantify their impact on performance, primarily measured by PSNR and SSIM on the LOL validation set. The following configurations are considered essential for this analysis:

1. **Baseline Model:** A student network trained with only the primary reconstruction loss (e.g., $\mathcal{L}_{\text{recon}}$) and without the resolution-progressive training strategy (i.e., trained directly at the target 256x256 resolution for an equivalent computational budget or number of image exposures). This will establish a fundamental performance baseline.

2. **Effect of Progressive Resolution Training (PT):** The baseline model augmented solely with the resolution-progressive training strategy (64px $\rightarrow$ 128px $\rightarrow$ 256px). This will isolate and quantify the benefits of the curriculum learning approach on convergence, stability, and final performance.

3. **Impact of Hybrid Fidelity Losses (HFL):** The PT-enabled model further augmented with the full set of fidelity-based losses ($\mathcal{L}_{\text{recon}}, \mathcal{L}_{\text{freq}}, \mathcal{L}_{\text{perc}}, \mathcal{L}_{\text{msssim}}, \mathcal{L}_{\text{grad}}$), but crucially without adversarial training or explicit contrastive distillation. This will demonstrate the collective benefit of the multi-objective fidelity guidance in improving various aspects of image quality.

4. **Contribution of Contrastive Distillation (CD):** The HFL+PT model augmented with the contrastive distillation loss ($\mathcal{L}_{\text{cont}}$) from the MIRNet-v2 teacher. This will quantify the specific impact of explicit feature-level knowledge transfer from the structural teacher on the student's performance.

5. **Contribution of Adversarial Training (GAN):** The CD+HFL+PT model (which represents our full proposed model) compared against a variant trained without the adversarial loss component ($\mathcal{L}_{\text{adv-G}} = 0$) and consequently no discriminator training. This will measure the influence of GAN-based training on perceptual quality and realism.

6. **Role of Adaptive Feature Stretch (AFS):** The full model trained without the AFS block to assess its specific contribution to enhancing early feature representations and its impact on the final enhancement quality and metrics.

7. **Role of Gradient-Guided Convolution (GGC):** The full model trained without the GGC block to evaluate its specific impact on edge preservation, structural detail rendition, and artifact suppression.

8. **Combined Effect of AFS and GGC:** The full model trained without both the AFS and GGC blocks to understand their synergistic or combined contribution to the overall performance.

For each ablation configuration, the student model will be trained under identical conditions regarding the dataset, total training epochs (250), optimizer settings, and hardware to ensure fair comparisons. We hypothesize that each major component—progressive training, the diverse terms in the hybrid loss (especially contrastive distillation and perceptual/structural losses), adversarial training, and the specialized architectural blocks (AFS, GGC)—will demonstrate a positive contribution towards the final performance and the balance achieved by the full "Compact Clarity" model. The results from these systematic studies will provide deeper insights into the framework's design, the importance of each element, and guide future refinements.

4.7 Discussion on Limitations

While our proposed "Compact Clarity" framework demonstrates a promising balance between computational efficiency and low-light image enhancement quality, certain limitations are acknowledged and warrant discussion:

- **Performance Gap with SOTA Teacher (MIRNet-v2):** Despite significant efficiency gains (our student generator being ˜2.66x smaller), a PSNR gap of approximately $2.34\,\mathrm{dB}$ persists when compared to the much larger MIRNet-v2 teacher model on the LOL dataset. While the SSIM scores are very competitive, indicating strong structural preservation, closing this PSNR gap entirely while maintaining a similar level of compactness remains a challenge. This suggests that some fine-grained detail restoration or sophisticated noise suppression capabilities inherent in the larger teacher model, arising from its greater parameterization and depth, might not be fully captured by the current student architecture and distillation strategy.

- **Complexity of the Hybrid Loss and Hyperparameter Tuning:** The framework employs a 7-component hybrid loss function. While this allows for multi-objective optimization targeting various aspects of image quality, the relative weighting of these components ($w_{rec}, w_{frq}, \ldots, w_{con}$) requires careful empirical tuning. The optimal balance of these weights may not be universally applicable across different datasets, varying low-light characteristics,

or different student architectures. Finding this ideal balance is an iterative, empirical process and adds a degree of complexity to the training setup.

- **Dependency on Teacher Quality and Appropriateness:** The effectiveness of any knowledge distillation process is inherently tied to the quality and appropriateness of the teacher model(s). If the teachers have inherent biases, limitations, or artifacts in their outputs/features, these might inadvertently be transferred to or negatively influence the student network. Our assumption is that MIRNet-v2 and Zero-DCE provide strong and complementary guidance, but alternative or imperfect teachers could yield different results.

- **Generalization to Diverse Unseen Conditions:** Training and validation were performed exclusively on the LOL dataset. While this dataset contains diverse indoor and outdoor scenes, the model's robustness and generalization capability to entirely unseen types of low-light conditions (e.g., extreme noise levels from different sensors, non-uniform illumination patterns not well-represented in LOL, or images from vastly different domains like medical or astronomical imaging) would require further extensive testing on a wider array of datasets.

- **Computational Cost of Training:** Although the student model is designed for efficient inference, the overall training process itself remains computationally intensive. This is attributable to several factors: the 250-epoch training duration, the resolution-progressive strategy (which involves data loading and processing for different image sizes), the inclusion of adversarial training (requiring updates to both generator and discriminator networks), and the feature extraction process for contrastive distillation (especially if teacher features are computed on-the-fly rather than pre-cached).

- **Ablation Study Status and Component Justification:** As explicitly noted, a full quantitative ablation study is planned as future work. Without these completed experiments, the precise individual contribution of each novel architectural component (e.g., AFS, GGC) and some specific loss terms (beyond the core reconstruction and perceptual losses) to the final performance is based on strong hypotheses and qualitative observations rather than rigorous empirical evidence from this specific study. The ablation studies are crucial for validating these design choices.

- **Interpretability of Distilled Knowledge:** While the framework aims to transfer knowledge, understanding precisely *what* knowledge is transferred and *how* the student utilizes it remains a general challenge in the

field of knowledge distillation, especially for complex tasks like image restoration.

Addressing these limitations could form the basis for future research endeavors, potentially involving more advanced distillation techniques, adaptive loss weighting schemes, automated hyperparameter optimization, further architectural refinements for the student network, or training on more diverse and challenging datasets.

Chapter 5
Conclusion and Future Scope

This thesis, titled "Compact Clarity: Development of Lightweight Networks for Low-Light Image Enhancement based on Multi-Objective Knowledge Distillation," embarked on the challenge of addressing the critical trade-off between enhancement performance and computational efficiency in the domain of low-light image processing. While large state-of-the-art models often achieve impressive visual results, their significant computational demands render them impractical for deployment in resource-constrained environments. Conversely, existing lightweight models may compromise on the quality and fidelity of the enhancement. Our research aimed to bridge this crucial gap by designing, implementing, and evaluating an efficient student network capable of delivering high-fidelity low-light image enhancement, achieved through an innovative progressive multi-teacher knowledge distillation framework. This chapter summarizes the core contributions and key findings of this work, discusses its broader significance, and outlines promising directions for future research.

5.1 Summary of Contributions and Findings

The research presented in this thesis has yielded several key contributions and findings, directly addressing the objectives outlined in Chapter 1:

1. **Novel Multi-Teacher Knowledge Distillation Framework:** We successfully designed and implemented a robust framework where a lightweight student network (Generator: 9.99M parameters; Generator + Discriminator for training: ˜12.76M parameters) learns from two expert teacher models possessing complementary strengths. MIRNet-v2 (˜26.6M parameters) provided guidance for structural and textural restoration, while Zero-DCE (˜0.1M parameters) offered conceptual expertise in efficient illumination adjustment. This multi-teacher strategy, particularly the explicit contrastive feature distillation from MIRNet-v2, allowed the student to inherit a balanced set of enhancement capabilities, outperforming simpler lightweight models.

2. **Effective Resolution-Progressive Training Strategy:** The introduction and application of a resolution-progressive training curriculum, systematically

advancing image sizes from 64x64 to 128x128, and finally to 256x256 pixels over 250 training epochs, proved highly effective. When coupled with dynamic batch size adjustment and gradient accumulation, this strategy facilitated stable convergence, allowed the network to learn coarse-to-fine features efficiently, and effectively managed GPU memory constraints during the demanding training process.

3. **Comprehensive Hybrid Loss Function for Multi-Objective Optimization:** A 7-component hybrid loss function was carefully formulated and successfully utilized to guide the student network's learning. This function incorporated terms for pixel-level reconstruction ($\mathcal{L}_{\text{recon}}$), frequency-domain fidelity ($\mathcal{L}_{\text{freq}}$), perceptual similarity ($\mathcal{L}_{\text{perc}}$), structural integrity ($\mathcal{L}_{\text{msssim}}$), gradient preservation ($\mathcal{L}_{\text{grad}}$), adversarial realism ($\mathcal{L}_{\text{adv-G}}$), and contrastive feature distillation ($\mathcal{L}_{\text{cont}}$). The meticulous weighting of these diverse components enabled the optimization for a wide range of image quality attributes, leading to results that are both visually pleasing and metrically strong.

4. **Efficient and Enhanced Student Architecture:** The student network, built upon an efficient backbone of Recursive Residual Groups (RRGs) and Multi-scale Residual Blocks (MRBs) (inspired by MIRNet-v2 but with reduced capacity), demonstrated its capability to learn complex enhancement mappings. The incorporation of optional enhancement blocks, namely Adaptive Feature Stretch (AFS) for dynamic feature range expansion and Gradient-Guided Convolution (GGC) for improved edge awareness, further augmented its feature representation capabilities without significantly increasing the parameter count. Activation checkpointing was also a crucial implementation detail for managing memory during the training of this architecture.

5. **Strong Performance on LOL Dataset with Significant Model Compression:** Extensive experiments on the standard LOL dataset validated the efficacy of our "Compact Clarity" framework. The proposed student model achieved a Peak Signal-to-Noise Ratio (PSNR) of $21.89\,\text{dB}$ and a Structural Similarity Index Measure (SSIM) of **0.858**. This performance significantly surpasses the lightweight Zero-DCE teacher and achieves a competitive SSIM score compared to the much larger MIRNet-v2 teacher. Notably, the student generator is approximately **2.66 times smaller** than the MIRNet-v2 generator configuration used, demonstrating a successful balance between performance and model compactness.

In essence, this research has successfully demonstrated that through a synergistic combination of multi-teacher knowledge distillation, a carefully designed and compact student architecture, a progressive training regimen, and a multi-

objective loss function, it is possible to develop lightweight networks that significantly advance the state-of-the-art in efficient low-light image enhancement. The objectives set forth at the beginning of this thesis have been substantially met, showcasing a practical and effective pathway towards deploying high-quality enhancement solutions in environments where computational resources are limited.

5.2 Significance of the Work

The primary significance of this work lies in its contribution towards making advanced low-light image enhancement capabilities more accessible and practical across a wider range of applications. By drastically reducing the parameter count and computational requirements compared to leading high-performance models, while still maintaining a high level of visual quality and structural fidelity, our "Compact Clarity" framework offers a viable solution with several implications:

- **Enabling Edge Device Deployment:** The reduced model size and improved efficiency facilitate the integration of sophisticated low-light enhancement algorithms into resource-constrained edge devices such as mobile phones, drones, IoT devices, and embedded camera systems, where processing power, memory, and battery life are at a premium.
- **Facilitating Real-Time or Near Real-Time Applications:** The lightweight nature of the proposed model moves closer to enabling real-time or near real-time enhancement for critical applications. This includes live video feeds for surveillance and security, improved night-time navigation aids in autonomous driving systems, and instant preview features in digital cameras and computational photography pipelines.
- **Democratizing High-Quality Enhancement:** By providing a model that balances quality with efficiency, this work helps make advanced image restoration techniques available to a broader range of users, developers, and applications that cannot afford the computational overhead or infrastructure requirements of larger SOTA models.
- **Methodological Advancement in Efficient Deep Learning:** This research showcases the power of combining progressive learning with multi-teacher, multi-objective knowledge distillation as an effective and structured strategy for model compression and efficient learning in complex low-level vision tasks. It provides insights into how different forms of knowledge (structural, illumination-based, perceptual) can be synergistically transferred to a compact student model.

- **Contribution to Sustainable AI Practices:** By focusing on smaller, more efficient models, this work aligns with the growing need for more sustainable AI practices that consume fewer computational resources both during training (though complex training is still a factor) and, more critically, during inference at scale.

This research pushes the envelope in the pursuit of efficient deep learning solutions that do not overly sacrifice performance, a critical direction as AI models continue to grow in complexity and are increasingly deployed in everyday technologies.

5.3 Future Scope and Directions

While this thesis has achieved its primary objectives, the field of efficient low-light image enhancement remains dynamic and offers numerous avenues for future exploration and improvement. Building upon the findings and limitations of this work, potential future research directions include:

- **Comprehensive Ablation Studies and Component Analysis:** As detailed in Chapter 4, conducting the planned ablation studies is a critical next step. This will precisely quantify the individual and combined contributions of each component (progressive training, specific loss terms, AFS, GGC, adversarial training, contrastive distillation) to the final performance, providing deeper insights for future model design and optimization.
- **Advanced Knowledge Distillation Techniques:** Exploring more sophisticated KD methods could yield further improvements. This might include distilling attention maps more directly from teachers, investigating relational KD to capture higher-order feature relationships, developing adaptive mechanisms to dynamically weight teacher contributions or loss terms based on training stage or input characteristics, or exploring data-free distillation methods.
- **Unsupervised and Self-Supervised Distillation Strategies:** Investigating methods to further reduce the reliance on paired ground truth data for the primary enhancement task is a valuable direction. This could involve leveraging teacher outputs in a fully unsupervised student training loop, incorporating self-supervised pre-training for the student network to learn robust initial features, or exploring cycle-consistent GAN approaches guided by teacher-defined quality metrics.
- **Extension to Low-Light Video Enhancement:** Adapting the current single-image framework to handle low-light video sequences presents a significant

and practical challenge. This would require incorporating temporal consistency constraints, leveraging temporal information from teacher models (if applicable for video), and potentially exploring recurrent or spatio-temporal student architectures to ensure smooth and coherent enhancement across frames.

- **Hardware-Aware Neural Architecture Search (NAS):** Utilizing NAS techniques, particularly those optimized for efficiency (e.g., targeting FLOPs, memory footprint, or latency on specific hardware platforms like mobile GPUs/NPUs), could automatically discover even more optimal lightweight student architectures specifically tailored for low-light enhancement on edge devices.

- **Robustness to Extreme Conditions and Diverse Noise Types:** Evaluating and improving the model's robustness to a wider variety of low-light conditions is important for real-world applicability. This includes performance under extreme underexposure, non-uniform illumination, diverse sensor noise profiles beyond those predominantly featured in the LOL dataset, and handling images with motion blur common in low-light captures.

- **Integration and Evaluation with Downstream Vision Tasks:** Assessing the practical impact of the enhanced images produced by "Compact Clarity" on the performance of downstream computer vision tasks (e.g., object detection, semantic segmentation, face recognition in low-light environments) would provide a more application-oriented evaluation of the enhancement quality beyond IQA metrics.

- **Exploring Alternative Teacher Architectures and Modalities:** As new SOTA models emerge (e.g., advanced Transformer-based architectures for image restoration, or models trained on different data modalities), investigating their potential as teachers to distill novel forms of knowledge or more robust representations into compact student networks is a continuous research avenue.

- **User Studies for Perceptual Evaluation:** Supplementing objective metrics (PSNR, SSIM) with subjective user studies involving human observers is crucial for a more holistic assessment of perceptual quality and visual appeal, especially when comparing subtle differences between enhancement methods or evaluating artifact presence.

- **Optimization for Specific Hardware and Deployment Constraints:** Further research into model quantization, pruning, and compilation techniques

specifically for the proposed architecture could lead to even greater efficiency and facilitate deployment on ultra-low-power microcontrollers or specialized AI chips.

By pursuing these and other innovative avenues, the research community can continue to develop increasingly powerful, efficient, and accessible solutions for overcoming the challenges posed by low-light imaging, ultimately benefiting a wide array of real-world applications and improving visual experiences in challenging environments.

List of Publications

This appendix lists the publications that have resulted from or are related to the research presented in this thesis.

1. JOSHI, KULADIPKUMAR And Bhat,Aruna, "Compact Clarity: Development of Lightweight Networks for Low-Light Enhancement via Multi-Objective Knowledge Distillation," Accepted for presentation at the *2025 International Conference on Computing Technologies & Data Communication (IEEE ICCTDC 2025)*, Hassan, Karnataka, India, July 4-5, 2025. (Paper ID: 2811; To be submitted to IEEE Xplore).

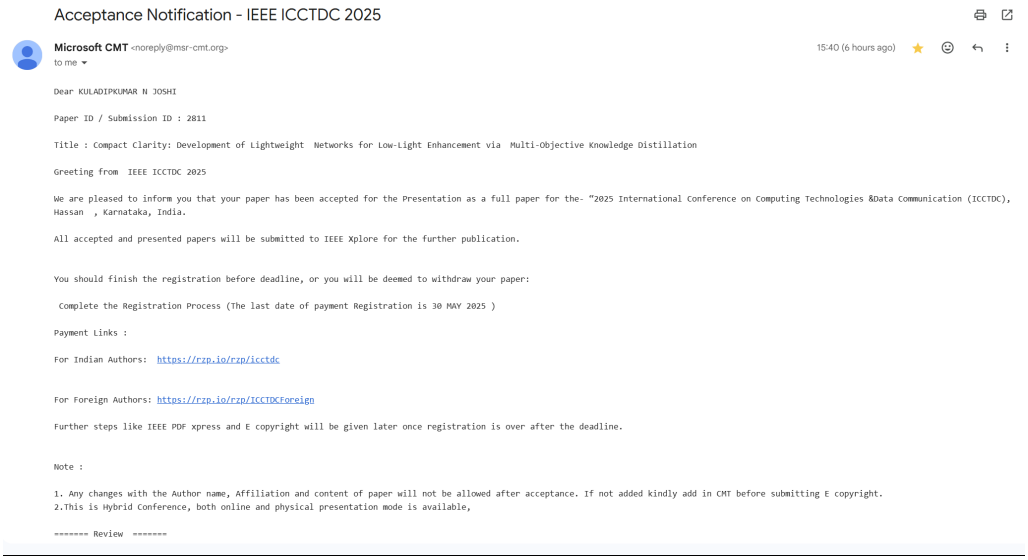
2. JOSHI, KULADIPKUMAR And Bhat,Aruna, "Illuminating the Darkness: A Comprehensive Survey and Future Directions in Low-Light Image Enhancement," In *Proceedings of the 6th International Conference on Inventive Computation Technologies (IEEE INCET 2025)*. Paper ID: 3164. Presented Online, May 22-23, 2025. (Certificate of Presentation Received).

Conference Certificates and Documentation

IEEE 6th INCET 2025

Figure 1: Certificate of Presentation for "Illuminating the Darkness: A Comprehensive Survey and Future Directions in Low-Light Image Enhancement" at the IEEE 6th International Conference on Inventive Computation Technologies (INCET 2025). Paper ID: 3164.

IEEE ICCTDC 2025 Acceptance



Acceptance Notification - IEEE ICCTDC 2025

Microsoft CMT <noreply@msr-cmt.org>
to me

15:40 (6 hours ago)

Dear KULADIPKUMAR N JOSHI

Paper ID / Submission ID : 2811

Title : Compact Clarity: Development of Lightweight Networks for Low-Light Enhancement via Multi-Objective Knowledge Distillation

Greeting from IEEE ICCTDC 2025

We are pleased to inform you that your paper has been accepted for the Presentation as a full paper for the- "2025 International Conference on Computing Technologies &Data Communication (ICCTDC), Hassan , Karnataka, India.

All accepted and presented papers will be submitted to IEEE Xplore for the further publication.

You should finish the registration before deadline, or you will be deemed to withdraw your paper:

Complete the Registration Process (The last date of payment Registration is 30 MAY 2025)

Payment Links :

For Indian Authors: https://rzp.io/rzp/icctdc

For Foreign Authors: https://rzp.io/rzp/ICCTDCForeign

Further steps like IEEE PDF xpress and E copyright will be given later once registration is over after the deadline.

Note :

1. Any changes with the Author name, Affiliation and content of paper will not be allowed after acceptance. If not added kindly add in CMT before submitting E copyright.
2.This is Hybrid Conference, both online and physical presentation mode is available,

====== Review ======

Figure 2: Acceptance Notification for "Compact Clarity: Development of Lightweight Networks for Low-Light Enhancement via Multi-Objective Knowledge Distillation" (Paper ID: 2811) for IEEE ICCTDC 2025.

Bibliography

[1] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1780–1789.

[2] J. A. Stark, "Adaptive image contrast enhancement using generalizations of histogram equalization," *IEEE Transactions on Image Processing*, vol. 9, no. 5, pp. 889–896, 2000.

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

[4] E. H. Land, "The retinex theory of color vision," *Scientific American*, vol. 237, no. 6, pp. 108–128, 1977.

[5] D. J. Jobson, Z.-u. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Transactions on Image Processing*, vol. 6, no. 7, pp. 965–976, 1997.

[6] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for fast image restoration and enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1980–1998, 2023.

[7] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. Pearson, 2018.

[8] D. J. Jobson, Z.-u. Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex," *IEEE Transactions on Image Processing*, vol. 6, no. 3, pp. 451–462, 1997.

[9] Z.-u. Rahman, D. J. Jobson, and G. A. Woodell, "Multi-scale retinex for color image enhancement," in *Proceedings of the International Conference on Image Processing (ICIP)*, vol. 3, 1997, pp. 1003–1006.

[10] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2011.

[11] X. Dong, Y. Pang, J. Wen, W. Xiao, and G. Wen, "Fast efficient algorithm for enhancement of low lighting video," in *2011 IEEE International Conference on Multimedia and Expo (ICME)*, 2011, pp. 1–6.

[12] S. Lee, J. S. Lee, H. S. Kim, and K. S. Kim, "Contrast enhancement using dominant brightness level analysis and adaptive intensity transformation for remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 4, no. 4, pp. 639–643, 2007.

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

[14] F. Lv, F. Lu, J. Wu, and C. Lim, "MBLLEN: Low-light image/video enhancement using squeezed-and-excited networks," in *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2018, pp. 1–6.

[15] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *Proceedings of the British Machine Vision Conference (BMVC)*, ser. BMVC 2018, 2018.

[16] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proceedings of the 27th ACM International Conference on Multimedia (ACMMM)*, 2019, pp. 1632–1640.

[17] Y. Zhang, J. Zhang, X. Guo, and J. Ma, "Learning to restore low-light images via decomposition-and-enhancement," in *Proceedings of the 29th ACM International Conference on Multimedia (ACMMM)*, 2021, pp. 2531–2539.

[18] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Transactions on Image Processing*, vol. 30, pp. 2340–2349, 2021.

[19] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *Computer Vision – ECCV 2020*, ser. Lecture Notes in Computer Science, vol. 12356. Springer International Publishing, 2020, pp. 492–511.

[20] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5728–5739.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008.

[22] C. Li, W. Yang, W. Ren, J. Liu, and J. Guo, "HWMNet: A hierarchical wavelet-based multi-scale network for low-light image enhancement," *IEEE Transactions on Image Processing*, vol. 30, pp. 9339–9352, 2021.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[24] X. Zhang, X. Zhou, M.-H. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6848–6856.

[25] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[26] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2015.

[27] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *International Conference on Learning Representations (ICLR)*, 2017.

[28] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4133–4141.

[29] S. Li, S. You, and H. Zhang, "Knowledge distillation with feature statistics matching," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.

[30] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3967–3976.

[31] Z. You, D. C. K. Chow, H. Zhao, M. Ye, F. X. Yu, P. Zhao, H. Jiang, and S.-F. Chang, "Learning from multiple teacher networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2017, pp. 1285–1294.

[32] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *International Conference on Learning Representations (ICLR)*, 2020.

[33] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[34] S. Ahn, S. Choi, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Computer Vision – ECCV 2018*, ser. Lecture Notes in Computer Science, vol. 11213. Springer International Publishing, 2018, pp. 252–268.

[35] T. Li, M. Liu, S. C. H. Hoi, and C. Wen, "Learning lightweight denoising network via knowledge distillation," *IEEE Signal Processing Letters*, vol. 27, pp. 1915–1919, 2020.

[36] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, 2009, pp. 41–48.

[37] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations (ICLR)*, 2018.

[38] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

[39] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3267–3275.

[40] T. Chen, B. Xu, C. Zhang, and C. Guestrin, "Training deep nets with sublinear memory cost," *arXiv preprint arXiv:1604.06174*, 2016.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014, published at ICLR 2015.

[42] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, 2003, pp. 1398–1402.

[43] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

[44] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[45] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.