BRIGHTNESS CONTROLLED IMAGE ENHANCEMENT DIFFUSION MODEL

A Thesis Submitted

In Partial Fulfillment of the Requirements
for the degree of

MASTER OF TECHNOLOGY

IN

ARTIFICIAL INTELLIGENCE

submitted by

NEHA THAKUR

2K22/AFI/12

Under the supervision of

Prof. Anil Singh Parihar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi 110042

MAY, 2024

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

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to Prof. Anil Singh Parihar for the continuous guidance and mentorship that he provided me during the project. He showed me the path to achieve my targets by explaining all the tasks to be done and explained to me the importance of this project as well as its industrial relevance. He was always ready to help me and clear my doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful.

Place: Delhi

Neha Thakur
(2K22/AFI/12)

Date: 31.05.2024

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

CANDIDATE'S DECLARATION

I, Neha Thakur(2K22/AFI/12), hereby certify that the work which is being presented in the thesis entitled "Brightness Controlled Image Enhancement Diffusion Model" in partial fulfillment of the requirement for the award of 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 January to May under the supervision of Prof. Anil Singh Parihar.

Neha Thakur
(2K22/AFI/12)

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

CERTIFICATE

Certified that Neha Thakur(Roll no 2K22/AFI/12) has carried out their project presented in this thesis entitled " Brightness Controlled Image Enhancement Diffusion Model" for the award of Master of Technology from the Department of Computer Science And Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis does not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/ Institution.

Date:

Prof. Anil Singh Parihar,
Professor,
Department Of Computer Science And Engineering,
Delhi Technological University,
Shahbad Daulatpur,Main Bawana Road,Delhi -42,

Abstract

Low-light image enhancement can be defined as the process of improving the visibility and quality of images captured in poor lighting conditions. The goal is to make these images clearer and more aesthetically appealing by increasing luminance, reducing noise, and refining details. However, many present algorithms focus on incrementing the brightness uniformly and to a particular extent which limits the experience of a user. This model utilizes a diffusion model that is conditioned on an illumination embedding. This framework allows the model to improve images in an iterative manner slowly refining the image quality by reducing noise and improving brightness. The illumination embedding serves as a control mechanism, enabling users to specify their desired brightness levels. This approach offers a high degree of adjustment which allows for tailored enhancements according to choice of the user. The embedding of the SAM gives freedom to users to choose particular areas in an image for the desired upgradation. This attribute improves user experience by providing an intuitive procedure for confined adjustments making the process efficient.

Contents

List of Tables

List of Figures

List of Abbreviations

CNN	Convolutional Neural Network
SAM	Segment Anything Model
SNR	Signal-to-Noise Ratio
GAN	Generative Adversarial Network
VAE	Variational Autoencoder
SSIM	Structural Similarity Index
PSNR	Peak Signal-to-Noise Ratio
DHE	Dynamic Histogram Equalisation
CLAHE	Contrast Limited Adaptive Histogram Equalisation
AHE	Adaptive Histogram Equalisation
CDF	Cumulative Distribution Function

Chapter 1

INTRODUCTION

1.1 Overview

A low-light image can be identified as a digital image taken under conditions where lighting is not sufficient to evidently illuminate the background. Some of the factors that attributes to this are: capturing images at night or evening time where natural light from the sun is minimal or absent, rooms or spaces that do have not good artificial light conditions, and weather conditions like foggy or stormy situations which can decrease the natural light, areas which are in profound shadows or concealed from direct light sources.

Some of the inherent characteristics of a low-light image are: high ISO camera settings or sensor impediments in low light contributing to the freckled or grainy image that is a noisy image, diminished information in both dark and light areas throws a challenge to capture a wide diversity of tones, color can look less muted or vibrant and image may lack contrast, blurriness in an image because of high exposure duration.

Enhancing a poor-lit image has many benefits like details are more comprehensible, clarity which is dominant for viewing, analyzing, and interpretation can make the image more appealing to the eyes, in security applications enhanced images are crucial for recognizing objects and activities, in the medical field it may direct to better diagnosis and treatment of disease, in research field it can help in analyzing occurrences which are not easy to observe, enhancing images of artifact and artwork contributes to better documentation and preservation, in content creation and social media enhanced images bring better results.

The scarcity of accessible paired datasets poses considerable issues in the development and training of productive image enhancement algorithms. The paired dataset comprises low-light images and their corresponding good-quality type which are used as ground truth references for training. Obtaining high-quality images demands guided lighting conditions and comparable camera settings which is an arduous task. We have to take pictures in such a way that we have to maintain the deviation between the paired images which is not easy.

In this project, we are using LOL [6] dataset. It consists of paired low-lit and well-lit images. It includes a total of 500 images, which are partitioned into four hundred and eighty five training pairs and fifteen testing pairs. It is obtained using a DSLR camera. It includes a wide mix of indoor and outdoor scenes which provides variety in textures and light conditions.

Denoising diffusion models [14] are getting recognition for image generation tasks due to their remarkable potential to generate high-quality images. It grasps the distribution of natural images by slowly adding noise to an image and then learning to reverse this process to generate an original image.

By iteratively denoising the image, it can produce a detailed image as they absorb the distribution better. It can generate a wide range of images that are diversified and realistic. The training process allows them to learn robust features. In comparison to earlier used methods like Convolutional Neural Networks and Generative Adversarial Networks, it demonstrates better performance, fewer artifacts, and greater stability.

Many existing approaches for the enhancement of low-light images are designed in such a manner that they follow injective transformation from dim-lit to well-lit images. This approach overlooks the inherently ambiguous nature where multiple possible well-lit versions of the same low light may exist. The model assumes that there is a uniform brightness level which makes a well-lit image. This approach lacks the flexibility to generate multiple possible enhancements that might be suitable for varied situations. This method has acute limitations in complex lighting situations where different regions of the image may be either under-exposed or over-exposed. This leads to sub-optimal results. This may worsen the over-exposed region, making them excessively bright and washed out.

So users can specify region of interest manually that is they can select a specific area of interest which may lead to localized adjustments instead of global, uniform change. It aims to improve the exposure in targeted regions, but accurate selection of regions requires precision and is cumbersome. On smartphones, people use their fingers to select regions on the touchscreen. Finger inputs are less precise than other methods. It may introduce noise and imprecise boundaries which makes the selection less accurate.

Instead, one can use unified illumination embedding to guide the enhancement of images. Here we use the diffusion model to enhance the image according to a target brightness level specified by the user. Illumination embedding represents the overall lighting condition of the image. It serves as a guide for the enhancement process. Here, the user can specify the desired brightness level which acts as a conditioning factor for the diffusion model. By, mentioning the desired brightness level they can control the final appearance of the enhanced image. Here, the brightness level is computed using the average pixel intensities of the image.

To further enhance the process we introduce the additional conditional elements, we condition the process with features derived from low-light images and supplement it with a normalized color map and SNR map. The diffusion process is conditioned on features procured from low-light images which aids in understanding the specific characteristics of an image. A normalized color map provides a standard reference for the colors in an image, ensuring the color enhancement is consistent. An SNR map indicates levels of noise in different regions. It helps to focus on areas with high noise which allows aggressive denoising wherever necessary. These supplements simplify the enhancement process which reduces the computational and complexity burden leading to faster convergence. The inclusion of these maps ensures that enhancement is not only focused on increasing brightness but also on maintaining color fidelity and reducing noise.

We also incorporate binary mask as additional input which facilitates localized edits. Segment Anything Model [17] is integrated in the framework to allow for the creation of a binary mask. It allows users to define regions of interest using simple and intuitive prompts such as points or boxes. This allows for user-friendly region control.

1.2 Motivation

Low light image improvement is an important area of research and practical applications. Enhancement improves visibility and extracts features that are otherwise hidden. Low -light introduces noise in the image which can be reduced by enhancement leading to clearer and aesthetically pleasing images. Even with rapid development in technology sensors struggles to capture good-quality images in bad lighting conditions. Enhancement of low-light image also has application in astronomy where celestial objects are faint. In cases like video calls or live-streaming, real-time low-light image enhancement improves the visual quality, making the experience better. Here, we use unified illumination embedding to enhance images. It captures the luminosity and lighting distribution of the image. Diffusion models have gained popularity in image restoration tasks. It works by learning the distribution of image by gradually adding the noise and reversing the process to obtain the image.

We use the diffusion model for low-light image enhancement. We further incorporate features derived from low-light images, along with a color map and SNR map. It makes the enhancement process becoming more effective. It ensures that low-light images are enhanced in such a manner that it addresses both the brightness and quality of the image, leading to optimal results.

1.3 Problem Statement

A low-light image is a photograph clicked under conditions of inadequate lighting. It can be encountered in various scenarios such as indoor setup with bad lighting conditions like dimly lit spaces, nighttime photography including landscapes, and underwater photography with diminished natural light. Enhancing low-light images involves methods aimed at boosting visibility, diminishing noise, and restoring details to produce visually pleasing and information-rich images.

The restraints of global enhancement methods become evident when we deal with images that contain both under-exposed and over-exposed regions. To address this issue user can specify region of interest manually to apply local edits. But, this manual process can be burdensome for smartphone users. Users must prudently delineate the area they want to enhance and meticulously trace boundaries.

So, we use a unified illumination embedding which represents the overall lighting of the image. The favorable results of denoising diffusion models have translated into image restoration tasks. Their capacity to model the natural distribution, iterative refinement, and adeptness makes them superior to many existing methodologies. To further amplify this process we use additional conditioning elements which eases the optimisation process. We condition the process with low-light image features, color map, SNR map. It results in a higher degree of improvement.

1.4 Research Objective and Contribution

The diffusion model architecture involves iteratively translating a simple distribution like Gaussian noise into a complex data distribution like natural image using a series of denoising steps.

A suitable choice for the neural network in diffusion models is the U-Net design. U-Net architecture is suited for image translation tasks owing to symmetric encoder and decoder structure along with skip connections. The encoder progressively downsamples the image, extracting high-level features. The decoder upsamples the encoded features, again creating the image and incorporating information from corresponding encoder layers through skip connections.

The noise schedule demonstrates how we add noise during the forward process and how it is removed during the reverse process. This is important for maintaining the stability and effectiveness of the model.

Here, we aim to enhance a low-light image using a denoising diffusion probabilistic model. We also condition the diffusion model with additional inputs like illumination embedding, normalized color map,SNR map, and binary masks. In this research, we enhanced and analyzed the model by following procedures:

U-Net Size: We increase the size of U-Net by adding convolutional layers to the model. By adding more convolutional layers IT can learn more intricate features, it can capture finer details and more nuanced patterns in data, leading to better efficiency.

Noise Schedule: The noise schedule is an important part that remarkably affects the performance and stability of the model. It defines how noise is iteratively added during each forward step and removed during the reverse denoising step.

Here, we analyze the model by adding different noise schedules and compare the performance of model for each schedule. We incorporate the following noise schedule in our model:

- Quadratic noise schedule
- WarmUp10 noise schedule
- Warmup50 noise schedule
- Linear schedule
- Constant noise schedule

Chapter 2

LITERATURE REVIEW

2.1 Traditional Light Enhancement Method

These procedures are used for single low light enhancement. A histogram of an image denotes the distribution of pixel magnitude values. Traditional histogram equalization is an image processing method used to intensify the contrast of a picture. It works by rearranging the pixel intensity values of a picture to cover a wider range, efficiently making dark regions lighter and vice versa. Traditional HE techniques, while effective in many cases, can sometimes result in over-enhancement, where image details are lost, or the introduction of unwanted artifacts.

DHE [1], introduces process for image contrast enhancement. Instead of applying histogram equalization uniformly across the entire image, DHE dynamically partitions the histogram based on the local features of the image. This partitioning is adaptive and varies depending on the image content. After partitioning, histogram equalization is applied to each sub-histogram individually. This process leads to more localized enhancement improving contrast in specific regions of the image without affecting the overall balance. An important factor of DHE is its ability to preserve the original brightness of the image.

CLAHE [20] presents an improved approach to traditional histogram equalization. Unlike global histogram equalization, AHE intensifies contrast in a local way within small regions of the image, making it more efficient for enhancing details in localized areas but it can amplify noise in homogeneous regions in an excessive manner. CLAHE introduces a contrast-limiting step to AHE. That is the histogram is clipped at a predefined value before computing the CDF hence limiting the amplification of noise.

[19] presents an approach for enhancing low-light images at the same time addressing the issue of noise. The limitations associated with low light image enhancement, mainly the trade-off between amplifying image details and compressing noise. Existing methods generally either enhance the image at the cost of increased noise or

denoise the image at the expense of losing important details. The aim is to propose a method that efficiently enhances low-light images while simultaneously denoising them, thereby improving both visibility and image quality.

Here, a two-stage approach is proposed. This technique is used to enhance the overall brightness and contrast of the low-light image. However, traditional AHE can introduce noise, so modifications are made to mitigate this issue. The gamma correction is applied to accomodate the brightness and contrast further. It makes the darker regions of the image brighter and amplifies the clarity of details.

After enhancing the image we use a bilateral filter to cut down the noise. Bilateral filtering is chosen because it can smooth the image while preserving edges, which is critical for maintaining important details in the image. In addition to bilateral filtering wavelet-based denoising techniques are used.

Retinex theory [18], presents a model to explain how humans perceive and discern colors consistently under changing lighting conditions. This model integrates both the physiological activity of the retina and the cognitive processes of the visual cortex.

In the initial stage retinal cells captures the light that is reflected from objects. This light is measured across different wavelengths corresponding to the primary colors red, green, and blue. The visual cortex of the brain processes these signals to interpret the final observed colors. This involves complex computations that consider the relative intensities of light in different parts of the visual field.

The image is decomposed into red, green, and blue components and is processed independently. For each pixel intensity of light is measured for each color channel. The intensity is compared for each pixel in a color channel to the intensities of surrounding pixels. This can be done using ratios or differences. The computed ratios are normalized to ensure they fit within a standard range. This step helps in maintaining the consistency of color perception. The processed color channels are then combined to form the final color-corrected image. This image should have consistent color perception across different lighting conditions.

2.2 Learning Type Light Enhancement Methods

Many learning based techniques for light improvement have been developed recently as a result of deep learning's speedy growth. These techniques enhance the grade of photos taken in low light by utilizing deep neural network feature extraction methods. They are especially helpful for difficult tasks like light enhancement because they can automatically learn to extract pertinent characteristics from raw image data without the need for human interaction. End-to-end training of deep learning models allows for optimization of the entire process, from input image to good quality image. Convolutional Neural Networks can manage spatial hierarchies in images, they can effectively capture both local and global properties that are essential for light enhancement. Multiple convolutional layers with different filters stacked on top of each other allows CNNs to gradually improve and refine picture attributes.

Supervised learning algorithms require paired datasets of low-light images and their corresponding high-quality images. The network learns mapping from low light images to enhanced images using training examples. Typical loss functions that aid in the model's ability to generate visually appealing outcomes are mean squared error (MSE) and perceptual loss.

For unsupervised learning algorithms, we use these when generally unpaired data is not available. We use methods like Generative Adversarial Networks. CycleGAN is an example of an unpaired technique that makes advantage of cycle consistency loss to guarantee that the enhanced image corresponds to the original low-light image when it is converted back to the low-light domain.

Some techniques optimize the picture enhancement process based on feedback from the environment by using reinforcement learning to learn policies.

A Deep Autoencoder Approach [12], is a deep autoencoder created especially to improve low-light photos by enhancing low-light photos while maintaining important information and reducing noise, it seeks to increase their visibility and quality. An encoder and a decoder are incorporated the deep autoencoder framework. While the decoder reconstructs the enhanced image from this representation, it squeezes the input image into a lower-dimensional representation, learning crucial information. In addition to incrementing brightness, the network is made to reduce noise, which is frequently seen in photos clicked in low light. The autoencoder's structure and training procedure provide this dual capacity. A dataset of photos in low light and their equivalent well-lit images is used to train the autoencoder. Reducing the difference between the network's augmented image and the well-lit ground truth image is the training goal. A loss function mean squared error (MSE) is used for this. In order for LLNet to be trained effectively, it needs training data that is diversified. It can be tedious to accumulate sizable datasets of matched high-quality and low-light photos. Instead, methods that mimic noise and low light levels are used to synthesize the training data.

This synthesis involves gamma correction. It mimics the many brightness levels found in pictures. The training dataset mimics low-light settings by incorporating photos with varying brightness levels through the application of random Gamma corrections. improves the model's capacity to generalize to many low-light situations and aids in its learning of how to handle a broad variety of brightness changes.

Deep Retinex [22] for low-light image enhancement, leverages the Retinex theory within a deep learning construct. The Retinex theory, states that colours are seen by the human eye by breaking down images into components of illumination and reflectance. The technique breaks down an image into its illumination (lighting circumstances) and reflectance (object intrinsic qualities) components by integrating the Retinex theory into a deep learning framework. To separate the illumination and reflectance components of a low-light image, a deep neural network is utilized. By processing the illumination individually and maintaining the reflectance, this decomposition helps in improving the image. The method optimizes the decomposition and enhancement processes at the same time since it is trained from start to finish. This guarantees that the network acquires the ability to generate improved photos that are true to the original scene and aesthetically pleasant.

Enlightengan [16] is a research paper that has an approach for enhancing low-light images using a Generative Adversarial Network without utilizing paired datasets. By employing this method low-light image enhancement through deep learning becomes more feasible for real-world scenarios when paired datasets are limited. By leveraging the power of GANs, it can produce appealing and high-quality enhanced images from low-light inputs. The method uses a cycle-consistency loss to guarantee the quality and consistency of the enhanced images. This means that when the enhanced image is converted back to the low-light domain, it should resemble the original low-light image. Adversarial training is used by this model to raise the realism and calibre of the augmented images. The generator generates images that are indistinguishable from enhanced photos, and the discriminator network is trained to differentiate between real enhanced images and images produced by the generator. Numerous benchmark datasets are used to assess this model, showing its efficacy in a range of low-light scenarios and image formats. It enables better flexibility and applicability in real-world scenarios where such datasets are not accessible, by doing away with the requirement for matched low-light and well-lit photos. The technique is appropriate for a broad range of applications, such as photography, surveillance, and any circumstance needing low-light picture augmentation. It may be used to many kinds of low-light photographs.

[11] presents a zero-reference deep learning framework for low-light image enhancement, that is it does not rely on reference images for training. It estimates pixel-wise enhancement curves which regulate the illumination of low-light images efficiently while maintaining important details and colors. It works without the need for paired datasets of enhanced and low-light images, in contrast to typical supervised algorithms. It makes the model simpler to implement. To calculate pixel-wise enhancement curves, which modify the brightness and contrast of each pixel in the image, the technique uses deep neural network. The input low-light photos are used to immediately learn these curves. The technique may handle various illumination conditions within that image more properly by adjusting and enhancing different portions

of the image by predicting enhancement curves at the pixel level. A lightweight convolutional neural network is at the heart of the model, and its purpose is to predict enhancement curves for every pixel in the image. The network outputs the enhancement curve parameters after receiving an image with low light levels as input. These curves are put into the input image, accommodating the illumination and contrast to generate the enhanced image. This model is checked on varied benchmark datasets, showcasing its effectiveness across diverse types of low-light image contents.

CERL [5] is a research paper that has a comprehensible approach for enhancing low-light images while considering realistic noise. It incorporates many upgrading ways to enhance image quality by at once enhancing light and diminishing noise. Light enhancement and noise reduction are combined into a single, coherent framework by this model. This devised plan ensures that improvements in light don't result in increased noise. To improve the resilience and efficacy of the enhancement process, the framework includes realistic noise models that faithfully capture the noise characteristics found in low-light photos. By increasing the contrast of low-light photos, it raises the visibility of details and raises the overall quality of the image. The framework mimics the noise characteristics commonly present in low-light photos by employing realistic noise models. The light Enhancement component targets to improve the image's brightness and contrast to make it more aesthetically pleasing. It separates undesirable artifacts and simultaneously reduces the noise in the image, preserving important features and textures. The strength is to combine these two modules' optimizations, guaranteeing that the final output image achieves a high degree of clarity and enhancement without adding undue noise.

RECORO [23]: is a research paper that proposes a procedure for enhancing low-light images with the extra capacity of user-controlled regional adjustments. It allows people to dictate regions of interest for target brightness using masks, improving the flexibility of the enhancement process. It offers precise control over how low-light photos are enhanced. With imperfect masks—rough sketches or broad strokes—users can designate portions of the image that require enhancement. The model then refines these masks to produce exact and high-quality light enhancement. This technique is especially helpful in situations where some areas of an image need to be processed more carefully than others, like in surveillance, medical imaging, or photography. Users can designate regions of interest with rough masks. The model precisely enhances the specified locations meanwhile maintaining the image quality by tuning these imperfect masks. The imprecise mask is handled by the refinement network to provide an exact, precise mask that defines the areas that need to be improved.

2.3 Challenges

A stable and generalizable mapping from low-light to enhanced images is challenging for the neural network to learn because of the inconsistent brightness levels throughout the training set. Since the network is always attempting to adjust to new brightness targets, it may find it difficult to converge to an answer that performs well for all images. To solve this issue unified illumination embedding can be used where the brightness level is represented using the average pixel intensities.

Chapter 3

METHODOLOGY

3.1 Diffusion Models

A family of machine learning generative models produces new data samples using a learned representation of the training data distribution. They are able to create new data that are absent in the training set as they have grasped the underlying patterns, structures, and distributions of the input data.

GANs need careful tuning and instability handling, but they are good for applications that demand realistic, high-quality image production. Diffusion models are for applications where quality and flexibility are important as they provide consistent training and varied sampling at the expense of a slower generation speed. Although they often produce images of inferior quality, VAEs offer a structured probabilistic framework with consistent training, making them appropriate for problems demanding explicit density modeling and latent space manipulations.

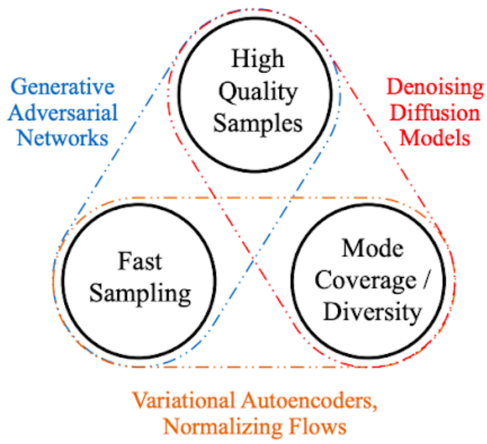

Figure 3.1: Comparison [10]

3.1.1 Workflow

Diffusion models came into inception due to disequilibrium thermodynamics. Here, we have steps to gradually add noise to data and then imbibe to reverse the method to create desired data from noise. It functions by dismantling training data via the consecutive addition of Gaussian noise and then understanding to recuperate the data by back-pedaling this process.

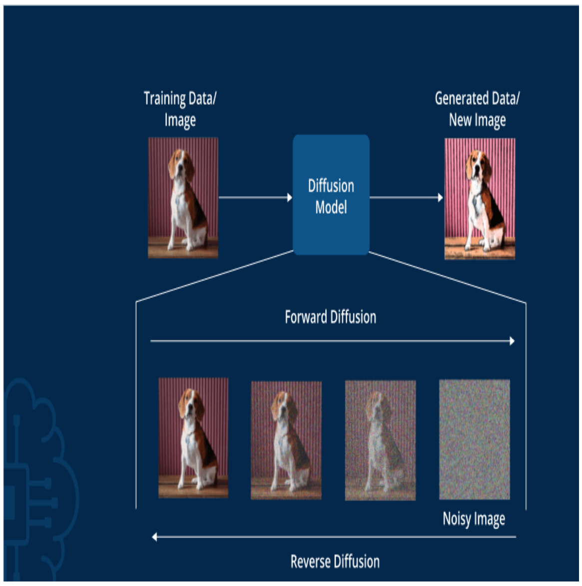

Figure 3.2: Working Of the Model [13]

It mainly comprises of two steps:

- **<u>Forward Process</u>**: The forward diffusion process is a method of slowly and iteratively adding noise in the data in a series of time steps, translating the sample into a noise distribution. In this process each step only depends on the previous state.

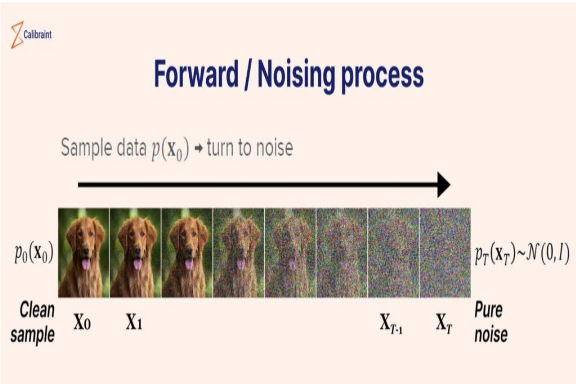

Figure 3.3: Forward Diffusion Process [3]

Let, $q(x_0)$ be the dispersal of the original images, then we can sample to get an image, $x_0$from $q(x_0)$.We can define the forward diffusion process $q(x_t|x_(t-1))$ which adds Gaussian noise at each time step t following to variance schedule β as:

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{3.1}$$

In closed form from x_0 to x_t can be reached in a tractable way. So,we can use reparameterization trick and equation(**??**) can be rewritten as:

$$q(x_t|x_0) = N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t I) \tag{3.2}$$

where $\alpha_t = 1-\beta_t$. The variance parameter can be fixed or singled out as a schedule over T time steps. Various types of variance schedules can be used like linear, cosine, and quadratic.

Figure 3.4: Variance Schedules [9]

- **<u>Reverse Process</u>**: The reversed process is the method in which the noise is progressively eliminated from the noisy data to produce samples that resemble the original data distribution. This process is essentially the opposite of the forward diffusion process and is used to produce new data points starting from pure noise.

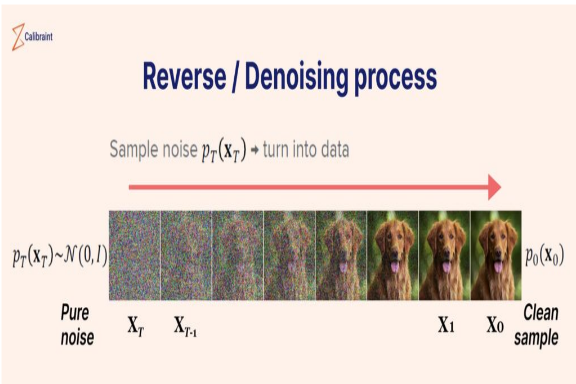

Figure 3.5: Reverse Process [4]

Mean has to be predicted as the variance is added according to the schedule. The final objective is to predict noise in the image between two timesteps. Mathematically,

$$p(x_{t-1}|x_t) = N(x_{t-1}; \mu(x_t,t), \Sigma_\theta(x_t,t)) \tag{3.3}$$

For loss function, negative likelihood is used: $-\log(p\theta)$.

$$-\log(p\theta) <= -\log(p\theta) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))) \tag{3.4}$$

Here KL divergence is added because it is being minimized. The above equation(**??**) is not computable either so some reformulation is done. After reformulation, our loss function reduces to:

$$L_{simple} = E_{t,x,\varepsilon}[||\varepsilon - \varepsilon_\theta(x_t,t)||^2] \tag{3.5}$$

ε_t is the pure noise generated at t.

3.1.2 Components Of Diffusion Model

Components of diffusion model comprises mainly two components-

- **U-Net:** It captures both local and global features of the data, the U-Net [21] design is a good choice for neural networks in diffusion models.

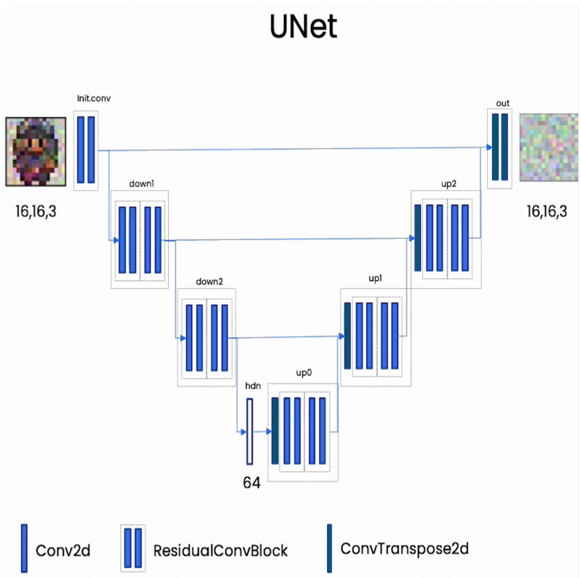

Figure 3.6: U-Net [7]

An encoder and a decoder make up the U-Net architecture. Skip links between the encoder and decoder directly link appropriate layers. Because of its architecture, the network can integrate high-level contextual data with low-level details, which is essential for jobs that need exact reconstructions, including the creation of images for diffusion models.

Encoder(Downsampler)-The encoder of a U-Net architecture is accountable for withdrawing features from the input data by progressively downsampling it via a sequence of convolutional neural network and max-pooling layers with help of skip connections. This allows the neural network to capture hierarchical features of the input, ranging from low-level details to high-level attributes. These stages are often referred to as blocks or layers.

Each block starts with convolutional layers which can be many. They put in convolutional filters to the input, producing feature maps that filter out aspects of the input data, such as edges and textures. The convolutional layers are usually followed by a non-linear activation functions.

Batch normalization layers can be appended after the convolutional layers to normalize the output, which aids in maintaining and speeding the training process.

Each block ends with a downsampling operation, typically max pooling. It decreases the structural dimensions of the feature maps while holding the most crucial part. This step increases the receptive field of subsequent layers, allowing the network to capture more contextual information.

As the input goes via each block of the encoder, the spatial dimensions of the feature maps decrease due to the pooling operations, while the depth increases which makes it possible for the network to capture and condense important spatial features.

An essential component of the encoder is skip connections, they directly transmit the output of each encoder block to their corresponding decoder block. This helps in preserving spatial features that might be lost during the downsampling process. It enables the network to combine low level structural information from the encoder with high level information in the decoder.

Decoder(Upsampler)-It is responsible for upsampling the feature maps produced by the encoder to generate output. It echoes the encoder's arrangement but in opposite manner, slowly incrementing the structural dimensions while diminishing the deepness of the feature maps.

Each block commences with an upsampling function to increase the spatial dimensions of the feature maps. Transposed convolution also known as deconvolution, is commonly used for upsampling. It understands to upsample the feature maps meanwhile learning missing information. Simple interpolation methods like bilinear or nearest-neighbor interpolation can be used for upsampling, followed by a regular convolution to filter the feature maps.

Each block mostly contains one or more convolutional layers to filter the upsampled feature maps. These layers apply convolutional filters to capture and intensify features, making sure that the produced output is of good quality and coordinates with the input data. This ceaseless process allows to recuperate spatial details which got disoriented while in the downsampling process in the encoder and attach low-level and high-level information from the skip connections for precise reformulation.

The decoder also uses skip connections to transfer feature maps from the encoder to the decoder. These connections make sure that the decoder has an approach to both low-level spatial information and high-level semantic information, facilitating precise reconstruction of the input data.

Timestep Embedding: In the U-Net architecture, timestep embeddings are important for conditioning the model on the particular diffusion step it is handling. These embeddings supply temporal context to the model, helping it comprehend the gradual increment of the diffusion process.

The diffusion process involves slowly adding noise to the data over many timesteps. During training and generation, the model wants to know the particular timestep it is dealing with to accurately predict and remove noise. Timestep embeddings encode this temporal information and integrate it into the neural network.

Timestep embeddings can be formed using sinusoidal functions, learned embeddings, or other methods. The common approach is sinusoidal embeddings, motivated by the positional encodings used in transformers. Sinusoidal embeddings encode the timesteps which is a combination of sine and cosine functions of different frequencies, allowing the model to grasp periodic patterns throughout timesteps.

The timestep embedding is concatenated to the input data or intermediate feature maps at a multitude of points in the network. This makes sure that the temporal context is considered throughout the processing. The embedding is processed through a fully connected layer to match the dimensionality of the feature maps before being integrated.

Attention Layers: Attention layers mechanisms are consolidated into the diffusion models to strengthen the ability of the model to apprehend long reliance and enhance the quality of the produced outputs. It assists the model to zoom in on the most appropriate components of the input data. This ability is crucial for reconstructing the intricate patterns in the data.

Model reflects on the importance of different pieces of the input when generating a piece of the output. It computes a weighted sum of the input features, wherein weights are demonstrated by the similarity between features. Query (Q): Represents the feature to be updated. Key (K): Represents the features to compare with. Value (V): Represents the features to be aggregated. The self-attention mechanism is computed as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.6}$$

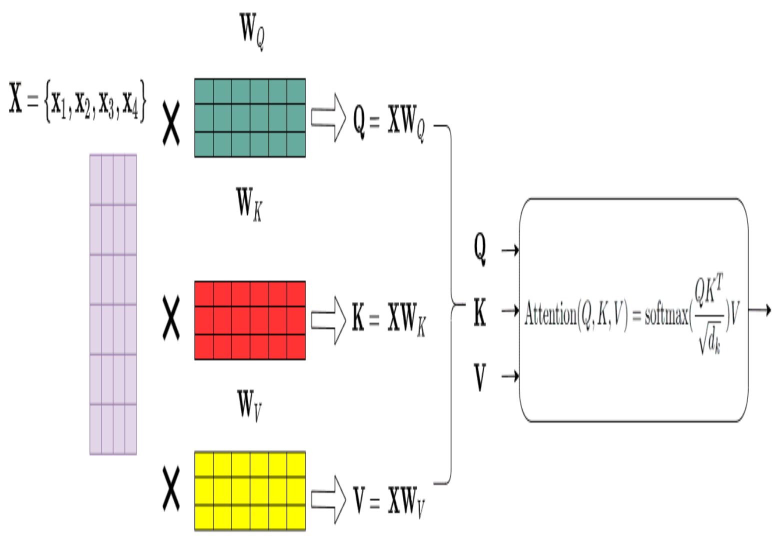

Figure 3.7: Attention [8]

In a U-Net used for diffusion models, attention layers are mainly appended at varied stages to improve the network's ability to deal with complex data. They can be placed in the bottleneck layer and in many places in both the encoder and decoder. It helps grasp dependence between distant parts of the input, which is crucial for generating coherent and detailed outputs. By focusing on important features, attention layers upgrades the quality of the learned representations, gearing towards better performance.

- **Noise Schedule:** In diffusion models, the noise schedule states how the noise is infused in the data with a sequence of timesteps. The choice of noise schedule is important because it affects the learning method and the grade of the created samples. There are many types of noise schedules:

  Linear Schedule: In a linear noise schedule, the noise level increments in a linear manner over time. It can be defined as:

$$\beta_t = \beta_{start} + t\left(\frac{\beta_{end} - \beta_{start}}{T}\right) \tag{3.7}$$

where β_t is the noise variance, β_{start} is the initial noise variance, β_{end} is the final noise variance, and T is the number of time steps.

Quadratic Noise Schedule: This noise schedule increments the noise variance in a quadratic fashion over time. It is defined as:

$$\beta_t = \beta_{start} + \left(1 - \cos\left(\frac{2\pi}{2T}\right)\right)(\beta_{end} - \beta_{start}) \tag{3.8}$$

where β_t is the noise variance, $\beta_s tart$ is the initial noise variance, $\beta_e nd$ is the final noise variance, and T is the count of time steps. It is inspired by the use of cosine annealing in learning rate schedules, providing a non-linear but smooth progression of noise.

Exponential Noise Schedule: An exponential noise schedule increments the noise level in exponential fashion over time. It is defined as:

$$\beta_t = \beta_{start} * \left(\frac{\beta_e nd}{\beta_s tart}\right)^{\left(\frac{t}{T}\right)} \tag{3.9}$$

where β_t is the noise variance, β_{start} is the initial noise variance, $\beta_e nd$ is the final noise variance, and T is the count of time steps.

Sigmoid Noise Schedule : It uses a sigmoid function to increment the noise level, which leads to a smooth transition that speeds up in the middle. It is defined as:

$$\beta_t = \beta_{start} + \left(\frac{1}{1 + (\exp -k(t - T/2)}\right)(\beta_{end} - \beta_{start}) \tag{3.10}$$

WarmUp10 Noise Schedule: It begins with a small noise variance and gradually increases it over a warmup period that is the first ten percent of the total timesteps. After this initial period, the noise variance increases more slowly. This helps the model stabilize in the early stages of training by avoiding large noise levels too soon.

WarmUp50 Noise Schedule: It works in a a similar pattern to Warmup10 but with a longer warmup period which is the first 50 percent of the total timesteps. It starts with a small noise variance and incrments more slowly over a longer period before transitioning to a more linear increase.

JSD (Jensen-Shannon Divergence) Noise Schedule: It is designed to minimize the Jensen-Shannon Divergence between the data distribution and the noise distribution over the diffusion process. It employs a logarithmic or other non-linear scaling of the noise variance to achieve a more balanced and grounded diffusion process. It can be defined as:

$$\beta_t = \beta_{start} \left(\frac{\log(1+t)}{\log(1-T)} \right) (\beta_{end} - \beta_{start}) + \beta_{start} \tag{3.11}$$

Constant Noise Schedule: It maintains a fixed noise variance throughout all timesteps. It can be defined as:

$$\beta_T = \beta_c onstant \tag{3.12}$$

Choosing the right noise schedule is crucial for the performance and stability of diffusion models, and it depends on the specific requirements of the task and data.

3.2 Addendum to the Model

1. **Color Map:**When enhancing low-light images, a common problem is the appearance of unnatural color shifts. It happens because the enhancement task can disproportionately amplify noise or color imbalances, which leads to color distortion that do not presents the true appearance of the locale.

   To handle this problem we can use a color map to normalize the range of the three color channels red, green, and blue in the input images. This normalization helps in preserving the balance between the color channels, and reducing color distortion, and keeping intact the natural appearance of the image.

   The three color channels are individually dealt with so that their values lie within a predicted normalized range. This is done so that no single channel commands the color balance of the image. This adjustment is crucial for maintaining the relative intensity levels across all channels.

   Let, an input image x be divided into three channels:

$$x = [x_r, x_g, x_b] \tag{3.13}$$

where x_r is red channel,x_g is green channel,x_b is blue channel.Let, the maximum pixel value for each channel be:

$$x_{max} = [x_{rmax}, x_{gmax}, x_{bmax}] \tag{3.14}$$

Then, the color map can be calculated as:

$$C(x) = \frac{x}{x_{max}} \tag{3.15}$$

The color map takes care that the process is uniform across all channels, thus maintaining the natural color relationships in the image.

2. **SNR Map:** Noise in images can notably deteriorate the grade of the image, it becomes difficult to discern details and true colors. Noise is more evidently present in low-light images. One such method is to use an SNR-aware transformer for effective enhancement.

   SNR is a metric that compares the level of the true signal of image content to the level of background noise. A higher value indicates clearer and superior-quality images with less noise. In low-light conditions, the SNR decreases because the signal is weak and the noise is more pertinent.

   An SNR map is designed to focus on regions of the image with low signal-to-noise ratios. This map mainly highlights areas where noise is more likely to be a trouble. The transformer applies the map for spatial attention. This means the model focuses more on regions with low SNR that is high noise and treat them differently compared to regions with high SNR that is low noise. By drawing attention to low-SNR regions, the transformer can allocate greater resources and apply much better noise reduction techniques in these areas. This targeted approach helps in reducing noise effectively. The capacity of the transformer to comprehend contextual relationships within the image makes it possible to enhance the image quality. The SNR map can be calculated as follows:

$$S(x) = \frac{F(x)}{|x - F(x) + \varepsilon|} \tag{3.16}$$

   where F is a low pass filter and ε is used for stability.

3. **Brightness Control Module:** Here, brightness is taken as a continuous class which means that instead of distinct categories, we represent brightness levels on a continuum. The vanilla brightness level lambda of an image is calculated by taking the average pixel value of a normal-lit image. This step makes sure that the brightness information is embedded in a way that maintains its properties also maintaining smooth interpolation. The illumination embedding, which now contains the encoded brightness information, is integrated into a U-net architecture using a Brightness Control Module.

4. **Region Controllability:** To address the need for judiciously increasing bright-

ness in particular regions of an image instead of globally illuminating the entire image, region controllability is incorporated.

A binary mask is used to describe the regions of interest in the image where brightness enhancement is the priority. The mask is a matrix of a similar size as the image with values of 1 in regions where enhancement is needed and 0 otherwise. It is appended with the original image inputs that allows the model to discriminate between regions that require brightness magnification and those which do not. By incorporating the mask [23], the model can zoom in on its enhancement efforts on the areas indicated by the mask.

To train the model effectively, synthetic data is generated using randomly sampled free-form masks having feathered boundaries. Feathered boundaries aid in the creation of smooth transitions between regions which are enhanced and not enhanced avoiding harsh edges. The synthetic target images are created by alpha blending low-light and normal-light images from existing low-light datasets [2]. Alpha blending is a technique where each pixel value of the target image is a weighted sum of the corresponding pixel values from the low-light and normal-light images, based on the mask.

3.3 Implementation

- We select a pair of low-lit and normal-lit images in random manner.
- We prepare supplemental information like a color map and SNR map, and then we generate a noisy version of the normal-light image using a forward diffusion process.
- Then we concatenate these maps and the low lit image which becomes the input to the model. Then we train the model to enhance the low-light image by leveraging the input, boosting both brightness and quality throughout maintaining color accuracy, and reducing noise.
- We treat brightness levels as ”class” labels for training a conditional diffusion model. Overall we train two models, a conditional [15] diffusion model and an unconditional model.

  The brightness level of an image is treated as a continuous class label. This lets the model handle varied brightness levels smoothly and adjust the brightness of images during the process. For conditional model:

$$\varepsilon_\theta(y_t, x, C(x), S(x)|\lambda) \tag{3.17}$$

where y_t is the image obtained after forward diffusion, x is the low lit image, C(x) is colour map, S(x) is SNR map, λ is the brightness level embedding. This model is learned to improve the image by conditioning on the brightness level λ. It uses the provided brightness level to adjust the enhancement process so that the output image meets the desired brightness.

For the Unconditional Model:

$$\varepsilon_\theta(y_t, x, C(x), S(x)|0) \tag{3.18}$$

where y_t is the image obtained after forward diffusion, x is the low light image, C(x) is the color map, and S(x) is the SNR map, Instead of a brightness level, a zero embedding (an array of zeros with the same shape as the brightness embedding) is used. The unconditional model is trained without conditioning on a specific brightness level.

- Users can enhance particular regions of an image by simply selecting the desired area with a click. This utilizes a binary mask to designate the regions for enhancement and combines the results from both the conditional and unconditional models. Users can click on the image to pick the region they want to enhance. This action creates a binary mask M, where the selected region is marked with ones and the rest of the image is marked with zeros. The binary mask is infused into the model inputs. This mask describes which parts of the image should be enhanced.

  For conditional model:

$$\varepsilon_\theta(y_t, x, C(x), S(x), M|\lambda) \tag{3.19}$$

  where y_t is the image obtained after forward diffusion, x is the low light image, C(x) is the color map, S(x) is the SNR map, λ is the brightness level embedding, M is the binary mask indicating the selected region. This model enhances the selected region based on the specified brightness level λ.

  For the Unconditional Model:

$$\varepsilon_\theta(y_t, x, C(x), S(x), M|0) \tag{3.20}$$

  where y_t is the image obtained after forward diffusion, x is the low light image, C(x) is the color map, S(x) is the SNR map, instead of a brightness level a zero embedding is used, and M is the binary mask. The unconditional model is trained without conditioning on a specific brightness level.

  The output procured from the conditional model, which targets the specified brightness level is used for the regions marked by the mask M. The output from the unconditional model is used for the rest of the image. The results from both models are mixed using the binary mask. It can be achieved by alpha blending, where the mask M controls the mixing of the two outputs. This means that only the selected regions are enhanced according to the specified brightness level, while the rest of the image receives a general enhancement.

Dataset: The LOL dataset comprises of five hundres low light and normal light image pairs which is sliced into four hundred eighty five training pairs and fifteen testing pairs. Here, low light images have noise generated during the photo-clicking process. Mostly the photos are captured in indoor background. All the photos are of 400×600 resolution.

3.4 Loss Function

- L_{simple}: With the reference to the loss function of the diffusion model (3.5) we concatenate the color map, SNR map, and a noisy image y_t generated using forward diffusion process.

  $$L_{simple} = E_{t,y,\varepsilon}[||\varepsilon - \varepsilon_\theta(y_t, t, x, C(x), S(x)||^2] \tag{3.21}$$

  where x is the low-light image, C(x) is the color map, S(x) is the SNR map, and y_t is the noised image produced using the diffusion process.

  While training the model might produce images with color distortions and unprecedented noise. It happens because the simple loss function might not sufficiently direct the model to focus on criteria like color fidelity and noise reduction. To deal with these issues we use auxiliary losses which provide additional guidance on the denoised images. These losses help the model to learn more efficiently by enforcing constraints that enhance color accuracy and noise reduction.

- Brightness Loss: The brightness loss aims to maintain the total brightness or luminance level of the enhanced images so that they are visually consistent with the original images in the aspect of brightness. The angular color loss focuses on maintaining the color distribution of the brightened images so that the colors remain in accordance with the original images.

- Angular Colour Loss: It measures the angular difference between the color vectors of the upgraded images and the ground truth images in a color space

- SSIM Loss: SSIM loss calculates the structural likeness in between the brightened images and the ground truth images considering both luminance, contrast, and structural information.
- Perceptual Loss: It aims to make sure that the enhanced images have similar perceptual qualities to the ground truth images by comparing high-level features extracted from both.

Chapter 4

RESULTS AND ANALYSIS

For low-light image enhancement, PSNR and SSIM are principal parameters to evaluate the standard of enhanced images comparison to the original ground truth images.

- **PSNR:** It guages the correlation between the largest potential of a signal in an image and the capacity of corrupting noise which impacts the correctness of its presentation. It is measured in decibels.

  For low light image upgardation PSNR is used to quantify how similar the enhanced image is to the reference image. A higher PSNR shows that the algorithm produces images that are closer to the reference in terms of pixel accuracy inferring better enhancement quality.

- **SSIM:** It measures the likeness between two images based on their luminance and shape. It is designed to simulate human visual perception more closely than PSNR. SSIM evaluates how nicely the enhanced image maintains the structural information of the reference image. An SSIM value near to one indicates high similarity demonstrating that the enhancement algorithm maintains the structural coherence of the image.

A blend of high PSNR and SSIM values indicates a successful enhancement algorithm.

U-Net: It enhances the ability to capture and rebuild complex image details while efficiently handling noise which is core to the success of diffusion models. Relative to the base paper CLE Diffusion [24], here we increase the size of the U-Net model by stacking convolutional layers in the head of the model. Here, we have performed 1500 epochs.

Model/Metrics	PSNR	SSIM	Average PSNR	Average SSIM
Base	23.124	0.778	23.352	0.794
Ours	24.5514	0.882	24.551	0.882

Table 4.1: Correlation

We can observe that by increasing the size of the U-Net accuracy of the model increases as PSNR value and SSIM value has increased.

Noise Schedule: It is a critical integrant that impacts the performance and productivity of the model. A noise schedule refers to the sequence and intensity of noise levels added to data during the forward process and removed during reverse process. We analyzed the performance of the model by integrating various noise schedules one by one. Here, we have performed 1500 epochs.

Schedule,Metrics	PSNR	SSIM	Average PSNR	Average SSIM
Linear	23.123	0.779	23.352	0.794
Quadratic	18.468	0.646	18.847	0.646
Constant	5.3136	0.0131	5.314	0.013
WarmUp50	17.695	0.374	17.695	0.374
WarmUp10	5.378	0.0135	5,378	0.0135

Table 4.2: Analysis Using Various Noise Schedules

We can observe that by applying different noise schedules there is variation in the calibre of the model. We get the best conduct using a linear schedule, then by quadratic schedule followed by the WarmUp50 schedule and constant schedule.

Visual Presentation of Results:

Base Model:

Epoch 0:

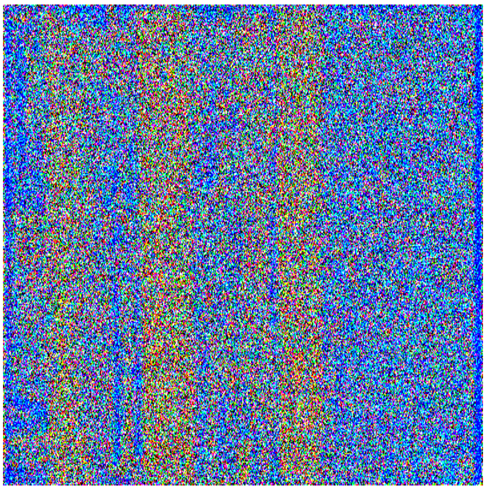

Figure 4.1: Image1

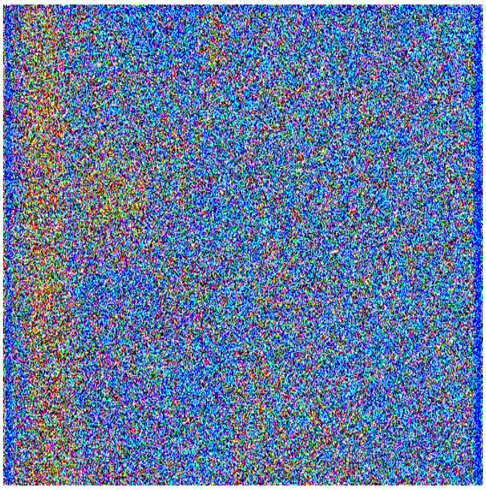

Figure 4.2: Image2

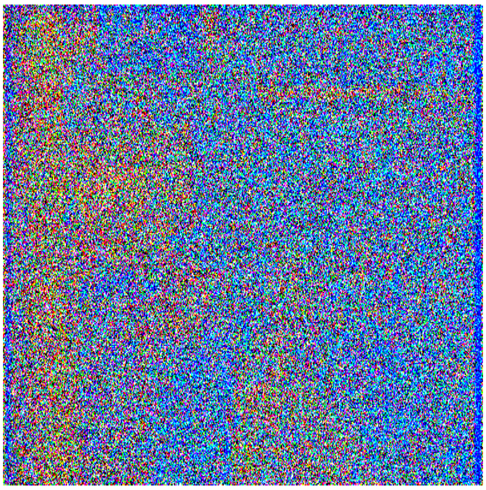

Figure 4.3: Image3

Epoch 1500:

Figure 4.4: Image4

Figure 4.5: Image5

Figure 4.6: Image6

Upgraded Model:

Epoch 0:

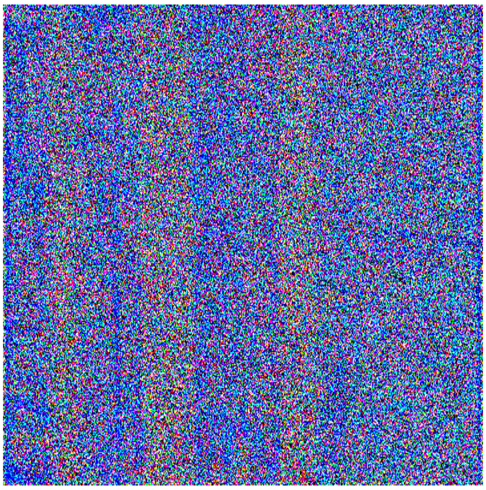

Figure 4.7: Image7

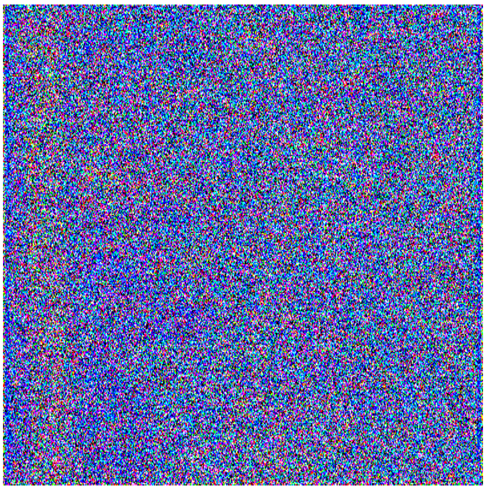

Figure 4.8: Image8

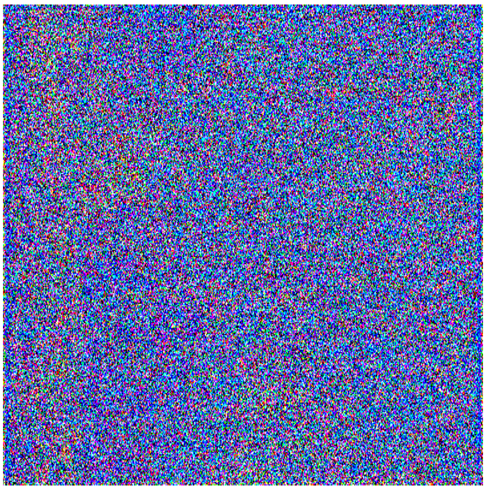

Figure 4.9: Image9

Epoch 1500:

Figure 4.10: Image10

Figure 4.11: Image11

Figure 4.12: Image12

Quadratic Noise Schedule:

Epoch 0:

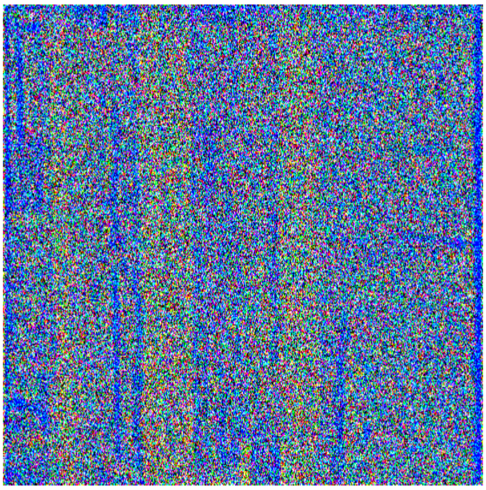

Figure 4.13: Image13

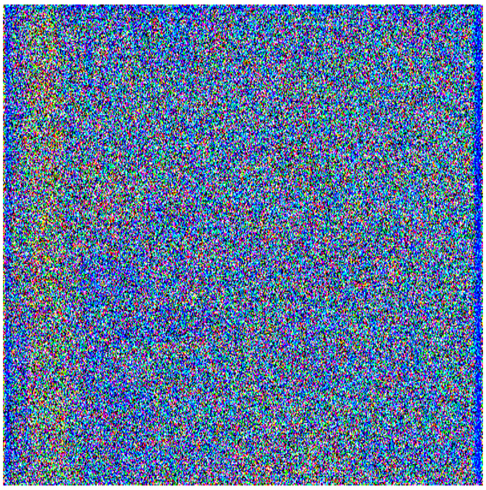

Figure 4.14: Image14

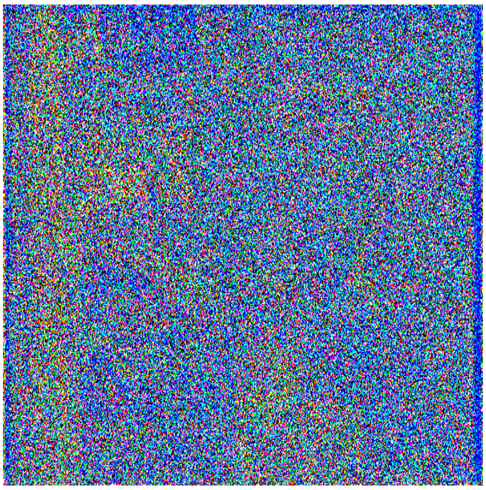

Figure 4.15: Image15

For 1500 epoch:

Figure 4.16: Image16

Figure 4.17: Image17

Figure 4.18: Image18

WarmUp 10:

Epoch 0:

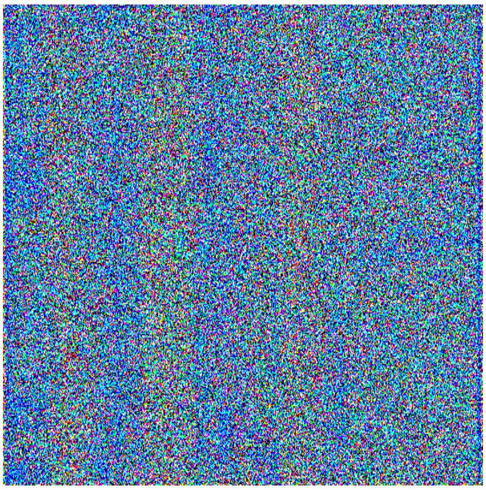

Figure 4.19: Image19

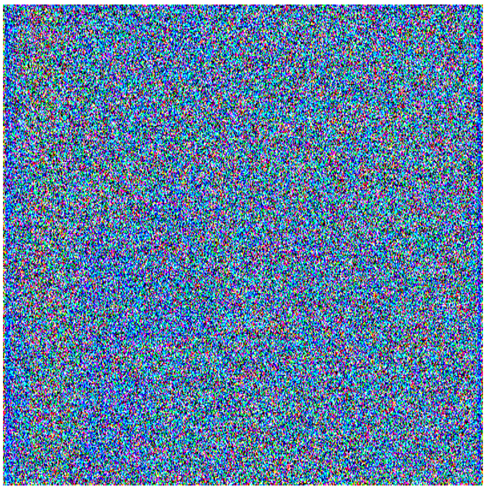

Figure 4.20: Image20

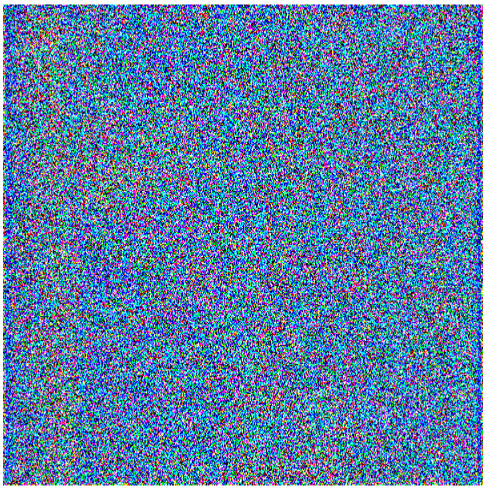

Figure 4.21: Image21

Epoch 1500:

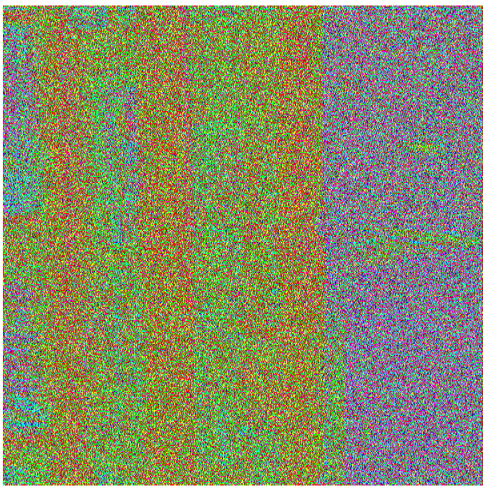

Figure 4.22: Image22

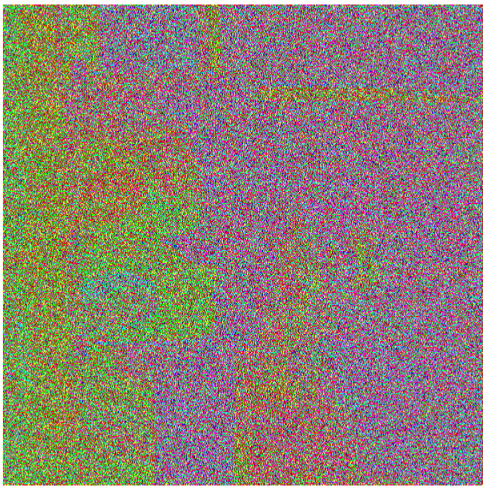

Figure 4.23: Image23

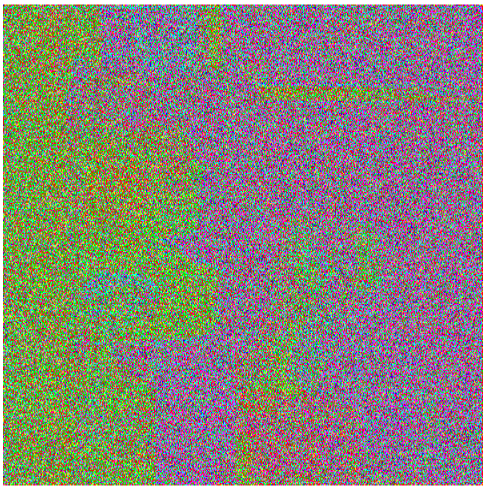

Figure 4.24: Image24

Linear Schedule:

Epoch 0:

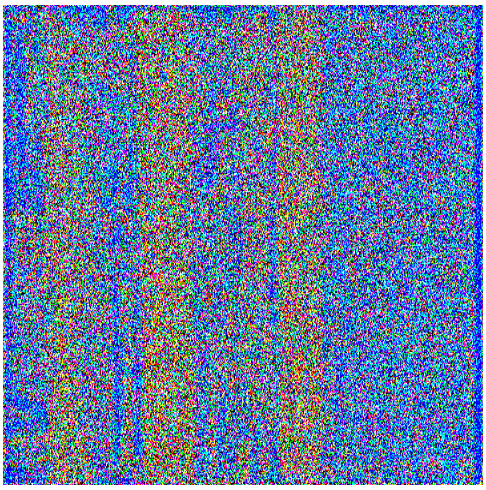

Figure 4.25: Image25

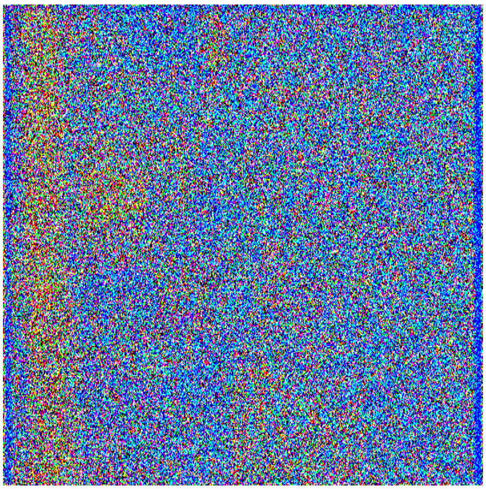

Figure 4.26: Image26

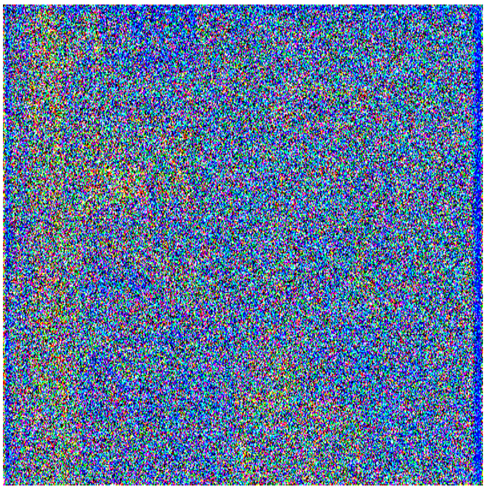

Figure 4.27: Image27

Epoch 1500:

Figure 4.28: Image28

Figure 4.29: Image29

Figure 4.30: Image30

Chapter 5

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

The model comprises a diffusion model prepared with an illumination embedding which gives precise check over the brightness of images in the inference stage. By conditioning the diffusion process on an illumination embedding, the framework allows for intuitive control of image brightness. The addition of SAM complements the usability of CLE by allowing users to choose particular regions of an image with a single click. This capability is powerful for targeted light enhancement, making the framework versatile and user-friendly.

We can extend the framework to other types of image enhancement tasks other than light control. We can further explore enhancements to the diffusion model and illumination embeddings for even better performance.

The social impact of the model can be substantial and multidimensional covering various domains from personal photography to professional media production and accessibility. We can extend the framework to other types of image enhancement tasks other than light control. We can further probe improvements to the diffusion model and illumination embeddings for even better performance. Using this model individuals can easily improve the lighting of their photos leading to higher-quality images for personal use and sharing on social media platforms. This can increase user satisfaction and engagement. For photographers, videographers, and media producers, it can simplify the post-production task by providing an efficient tool for enhancing lighting in images and videos. This can save one from manual editing.

The model can help maintain persistent lighting across a series of images or video frames, which is important for professional media production and keeping a high standard of visual quality. By enhancing the lighting in images it can make visual content more approachable to people with low vision helping them see details more clearly and enjoy visual media more fully. Educators can use models to enhance the quality of visual substance allowing them to be clearer and more engaging for students.

References

[1] Mohammad Abdullah-Al-Wadud, Md Hasanul Kabir, M Ali Akber Dewan, and Oksam Chae. A dynamic histogram equalization for image contrast enhancement. *IEEE transactions on consumer electronics*, 53(2):593–600, 2007.

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. IEEE, 2011.

[3] Calibraint. Forward process. https://www.calibraint.com/blog/beginners-guide-to-diffusion-models, 2023. Accessed: 2024-05-30.

[4] Calibraint. Reverse process. https://www.calibraint.com/blog/beginners-guide-to-diffusion-models, 2023. Accessed: 2024-05-30.

[5] Zeyuan Chen, Yifan Jiang, Dong Liu, and Zhangyang Wang. Cerl: A unified optimization framework for light enhancement with realistic noise. *IEEE Transactions on Image Processing*, 31:4162–4172, 2022.

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

[7] Riya Chikkara. U-net. https://www.linkedin.com/pulse/14-coding-u-net-architecture-from-scratch-riya-chhikara-xbvte/, 2024. Accessed: 2024-05-30.

[8] By DAIVI. Self-attention. https://www.projectpro.io/article/transformers-architecture/840, 2024. Accessed: 2024-05-30.

[9] By Deci. Variance schedules. https://deci.ai/blog/decidiffusion-1-0-3x-faster-than-stable-diffusion-same-quality/, 2023. Accessed: 2024-05-30.

[10] Ainur Gainetdinov. Strengths and weaknesses of models. https://pub.towardsai.net/diffusion-models-vs-gans-vs-vaes-comparison-of-deep-generative-models-67ab93e0d9ae, 2023. Accessed: 2024-05-30.

[11] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light

image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1780–1789, 2020.

[12] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *arXiv e-prints*, pages arXiv–1511, 2015.

[13] Leeway Hertz. Workflow. https://www.leewayhertz.com/diffusion-models/, 2024. Accessed: 2024-05-30.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[18] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977.

[19] Lin Li, Ronggang Wang, Wenmin Wang, and Wen Gao. A low-light image enhancement method for both denoising and contrast enlarging. In *2015 IEEE international conference on image processing (ICIP)*, pages 3730–3734. IEEE, 2015.

[20] Stephen M Pizer. Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen m. pizer, r. eugene johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group. In *Proceedings of the first conference on visualization in biomedical computing, Atlanta, Georgia*, volume 337, page 1, 1990.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference,*

Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015.

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

[23] Dejia Xu, Hayk Poghosyan, Shant Navasardyan, Yifan Jiang, Humphrey Shi, and Zhangyang Wang. Recoro: Re gion-co ntrollable ro bust light enhancement with user-specified imprecise masks. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1376–1386, 2022.

[24] Yuyang Yin, Dejia Xu, Chuangchuang Tan, Ping Liu, Yao Zhao, and Yunchao Wei. Cle diffusion: Controllable light enhancement diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8145–8156, 2023.