Zero Reference Low-Light Image Enhancement with Attention

Tamal Mondal
cs21mtech12001@iith.ac.in

Kamal Shrestha
cs21mtech16001@iith.ac.in

Aman Agrawal
cs21mtech11006@iith.ac.in

Praveen Vishwakarma
cs21mtech11010@iith.ac.in

Jayamohan.C.B
cs21mtech13001@iith.ac.in

Abstract

It is common to have images taken in low-light conditions due to environmental conditions like the extreme backlight or technical issues like camera sensing capacity. These images suffer from poor visibility, and on the other side, it does not convey complete information. Good quality images or videos are crucial for surveillance, autonomous driving, etc. Many image enhancement techniques have been developed using different approaches ranging from histogram equalization methods to machine learning methods. As a partial fulfillment for Deep Learning, AI5100, we took up a low-light image enhancement task using a deep learning-based method called Zero-Reference Deep Curve Estimation(Zero-DCE). The idea is to use carefully formulated non-reference loss functions to convert the light enhancement as an image-specific curve estimation task. The model is also lightweight, requiring limited computational resources for training and inference time, making it suitable for real-world applications. This project presents a proposed Convolution Block Attention Module (CBAM) approach over Zero-DCE Architecture with channel and spatial gate attention to generate better-enhanced images. It achieved a training loss of 0.7198 *with an improvement from Zero-DCE with a training loss of* 0.9639 *over* 10 *epochs. With the expense of a significant increase in inference time, the proposed model was able to generate better-enhanced images under different low-light conditions.*

1. Introduction

As already mentioned, having well-illuminated images are critical for various real-world image or video processing applications. Nowadays, with more and more cameras(particularly in smartphones), image samples are obtained under different lighting conditions such as night, extreme back-light, etc. The images captured under low light conditions often come as noisy, under-exposed, color distorted forms. These issues are particularly evident in phone cameras, where the sensors are less sensitive in low light conditions. The quality of the image captured by the camera depends on the sensor's properties such as aperture, exposure, shutter speed, etc. Some high-end cameras contain extra hardware for noise correction and illumination, but then it becomes costlier as more hardware gets included in the camera sensor suite.

By adjusting camera settings, the quality of the image can be enhanced to some extent. But, in a highly dark environment, generated images are often noisy. There are other issues also like when the exposure is long, it introduces blur, and on the other hand, if the exposure is short, that introduces noise. Typically, in low-light images, the signal-to-noise ratio(SNR) and photon count are low, which makes it a challenging problem. Increasing the ISO can be one way to increase the brightness, but that doesn't solve the problem, as it also increases noise. There can be other physical ways to increase SNR like increasing exposure time, using additional lights(ex. flash of mobile), etc. Still, these are not always feasible and have their own issues, like using flash can introduce more light than needed. So there is a real need for low-light image enhancement software so that we don't need to buy or carry good quality cameras if not necessary, and systems like autonomous cars, satellites, surveillance systems, etc., don't need to deploy additional hardware.

For this work, we have considered Zero-Reference Deep Curve Estimation(Zero-DCE) [5] as our baseline for the low-light image enhancement task. Interestingly, Zero-DCE doesn't need any paired or unpaired images during training, unlike CNN based [20, 24] or GAN based [8, 28] approaches, which helps to avoid over-fitting. This is possible because the loss function doesn't contain ground truth references. Zero-DCE models the image enhancement task as an image-specific curve estimation problem. The loss function combines exposure control loss, spatial consistency loss, illumination smoothness loss, and color constancy loss. The model takes a low-light

image as input and produces higher-order curves applied for pixel-wise adjustment to obtain an enhanced image. Additionally, the Zero-DCE model is lightweight (having only seven convolution layers) and has around 80k trainable parameters, which is relatively less for deep learning models. Due to this, Zero-DCE model training takes only 30 mins in an Nvidia 2080Ti GPU. During inference for 32 images of size $1200 \times 900 \times 3$, Zero-DCE takes 2.5 msec to process, which is way better with respect to its peers, and it makes Zero-DCE model suitable for real-time applications. We'll be looking at the baseline of Zero-DCE, its drawbacks compared to some of the more recent works and try to improve upon the existing model.

2. Literature Review

Illumination enhancement [13] has been one of the commonly known problems for a long time in the image processing domain. There have been many attempts to improve the quality of low-light images. A variety of techniques have been developed, starting from histogram-based methods to current deep learning-based approaches.

An experiment-based review of Low-Light Image Enhancement methods was carried out in 2020 [22]. It attempted to categorize various enhancement techniques based on the approaches. The methods are as follows:

- **Gray transformation methods**: These methods transform the gray values of single pixels into other gray values through a mathematical function.
- **Histogram equalization methods**: These methods use the cumulative distribution function (CDF) to adjust the gray output levels.
- **Retinex methods** - These methods are based on the illumination-reflection model; an image can be expressed as the product of a reflection component and an illumination component.
- **Frequency-domain methods** - These methods are based on frequency domain. It transforms an image into the frequency domain, performs an operation, and converts it back to the spatial domain.
- **Image fusion methods** - These methods take multiple images from the same source or multiple sources and fuse them to obtain an enhanced image.
- **Defogging model methods** - These methods are based on improving foggy images. It negates the image first and assumes it as a foggy image, and tries to remove such effects from the image.
- **Machine learning methods** - These methods use various types of machine learning techniques to solve the low-light image enhancement task.

Not only these, due to the recent advancements in the area of deep learning, many low light image enhancement algorithms have been developed that introduce a variety of techniques using different learning strategies ranging from supervised to adversarial leanings [11].

The initial attempts to generate an image to image mappings were using histogram equalization(HE) [1]. This method compared and tried to maintain the mean brightness of the original image inside an adjusted image, adjusted in terms of saturation, details(sharpness), and more.

To make the method more dynamic and usable in consumer electronics, we have [7], which focuses on maintaining the intensity of the original image with the resulting image by smoothing the intensity histograms using Gaussian filters followed by partitioning and assigning them into a dynamic range. Each range was then equalized and normalized individually using the histogram equalization process to maintain the mean intensity.

Following the above methods, we now have a different approach to tackle the low light image enhancement task. Retinex [10], and Retinex-Net [24] treat the enhancement of image as an adversarial learning task, where each image is factored into the product combination of reflectance and illumination. Each illumination component was used to formulate and learn an image-specific illumination map used to enhance consistent reflectance parameters between each pair of images.

However, the naturalness of non-uniform illumination images cannot be fully preserved since methods described in [10, 24] do not limit the range of reflectance and do not exclude illumination as the default choice. As in Retinex, a recommended bright-pass filter splits the image into reflectance and illumination. A novel bi-log transformation was implemented to map the illumination to establish a balance between details and naturalness of the image, where the realistic nature was measured using a novel lightness-order metric.

In continuation to the above methods, [4] proposed a weighted variational approach to estimate the reflectance and illumination from the input image.

Instead of getting the illumination map from the observed image, the natural illumination was calculated using the maximum intensity of each pixel in the previously enhanced RGB channel using the image structure. The

maximum values for the R, G, and B channels were used to estimate the illumination of each pixel. Structural priorities were imposed to obtain the final lighting map [6].

In [21], an optimization problem-based technique was employed to estimate the lighting map using noise. Instead of utilizing a logarithmic transformation, this method analyses structural information in low-light photographs to forecast the noise map, which is then solved using a Lagrange multiplier based alternating direction minimization algorithm. Previous methods unintendedly affected the image histogram distribution or relied on potentially incorrect physical models.

With the advancement of deep learning techniques, attempts were carried out in low-light imaging areas. In [11], a survey of deep learning-based approaches was carried out.

Convolution Neural Networks (CNN)s based Methods use paired images (low/normal conditions) or Paired Supervision. One of the drawbacks of learning-based approaches for the generation of training data. It is often a tedious and expensive task to obtain normal light and low light image pairs. There were attempts in the zero-shot learning strategy, where paired images are not required for training.

In [20], Instead of carrying out an image to image mapping, a mapping between the image to illumination was done, and the final output is generated from the learned illumination map. Here, the model learns with the use of intermediate lighting to achieve improved results with the supervision of complex photo adjustments retouched input/output image pairs.

We also have the generative adversarial approaches to enhance low light images using unpaired images. Photographing in low and normal light simultaneously is becoming increasingly difficult. [8] describes an unsupervised generative adversarial network(GAN) that can be trained without low light image pairs by learning to regularise unpaired training data using information extracted from the input, which includes a global-local discriminator structure, a self-regularized perceptual loss function, and the attention mechanism.

Similarly, LLNet [14] is a deep autoencoder-based approach. It identifies signal features from low-light images and adaptively brightens images without over-amplifying the lighter parts in images (i.e., without saturation of image pixels) in a high dynamic range.

Some of the more recent works include the use of a normalizing flow model. [23] focused on considering the complex conditional distribution of normally exposed images by training an invertible network to map the conditional distributions into Gaussian distributions. The model showed significant results in the LOL dataset [24] and is currently the best model for that dataset according to the average PSNR and SSIM metrics.

A paper focusing on optimizing the computational speed of the low-light enhancement of images was given by [9]. It preprocesses the images in higher scale-spaces and simultaneously enables processing in all scale-spaces. It also provides an "off-the-shelf" amplification module for pre-amplifying images requiring almost no fine-tuning. These optimizations give the result much faster while still being comparable to the other models in terms of metrics.

HWMNet [3] is an M-Net [16] based CNN-model that utilizes the hierarchical structure of the M-Net+ model (used for medical purposes) and adds a half-wavelet attention block to get enriched image features. It also gives a competitive state-of-the-art performance in the LOL [24] dataset.

Multi-Axis MLP for Image Processing (MAXIM) [18] provides a spatially-gated multi-layer perceptron (MLP) model for various image processing tasks and manages to achieve state-of-the-art performance in a lot of them, including deblurring and low-light image enhancement. It primarily has two building blocks that take care of spatial mixing of global and local visual spaces and cross-feature mutual conditioning, respectively, to enhance low-light images.

A multi-branch CNN guided by attention blocks was employed in [15], in order to perform brightness enhancement and denoising with their respective attention maps. They also synthesise a diverse dataset from publicly available datasets which helps with the model adaptation.

ABSGN [2] for low-light enhancement utilises a self guided network combined with global as well as channel attention blocks in a top-down approach in order to enhance the images. To exploit information extracted at multiple scales (resolutions), wavelet transformation is used to convert the feature maps. At the lower resolutions, it uses a Global Spatial Attention block to learn the global context and colour information, whereas in the higher resolutions it uses channel attention blocks to learn the attention weights for different channels. The paper claims to get better evaluation metrics in the LOL dataset than the previously mentioned results with a relatively low inference time.

Another highly effective image enhancement method

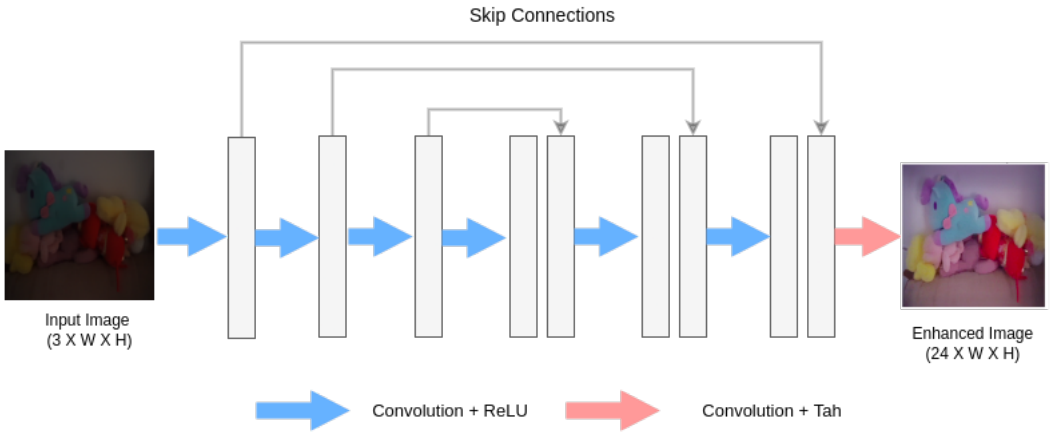

Figure 1. DCE-Net Architecture

that has been introduced recently is MIRnet [26]. What makes this technique so special is the multi scale approach in which only the unwanted degraded details from the image are discarded and at the same time useful spatial details are preserved. This model maintains the high resolution details hence preserving the spatial details to the maximum possible extent and simultaneously process lower spatial resolutions using parallel convolution streams. The key difference between MIRnet and other existing multi scale methods is that other methods process individual scales separately and in an isolated manner but MIRnet shares information across all scales and that too for all resolution levels.

Finally, we have Zero-Reference Deep Curve Estimation (Zero-DCE), which focuses on the image specific curve estimation to produce higher order curves from low light input images. Zero reference means that it doesn't require any paired or unpaired images of low light and adequate light images for the training process, like in CNN and GAN-based methods. Instead, this method works based on four loss functions that do not require any ground truth reference for training and efficiently measure the enhancement quality of the image. So, despite the lightweight nature of the model, the estimated image-specific curve generalizes well to diverse conditions and makes dynamic adjustments to low light input images.

Later, a faster and lighter version of Zero-DCE was also released, called as Zero-DCE++ [12] which had eight times lesser the number of parameters than its predecessor, translating to twice the run time speedup. It had significant improvements in training time as well without compromising much on the quality of enhanced images, thus proving effective for resource-limited devices.

3. Methodology

3.1. Zero-DCE

Zero-DCE is a method to enhance low light images into recognizable well illuminated ones using deep curve estimation specific to the input image. Zero-DCE method does not carry out image to image mapping and it does not require any labeled or unlabeled data to train with. Hence this method is called zero reference. Zero-DCE framework solves the image enhancement problem by estimating best fitting light enhancement curve from the input image. Light enhancement curves are estimated using simple 7 layer CNN, DCE-Net. All pixels in the input image is mapped to output image by repeatedly applying light image enhancement(LE) curves. LE curves are similar to curves available in image editing softwares. By changing the shape of the curve, intensity of pixel will be changed. LE curve's shape is controlled by weight parameter and it is estimated by CNN. Pixel wise LE curve parameter is estimated using DCE-Net. LE Curve map is calculated for each color channel of the input image and it is repeatedly applied to input image to get the enhanced output image. Light enhancement curves are estimated with the help of carefully designed loss functions. The LE curve estimation equation is as follows:

$$LE(I(x);\alpha) = I(x) + \alpha I(x)(1 - I(x))$$

$I(x)$ is the input image pixel intensity, α is the learn-able parameter of the LE curve. Value of α ranges from -1 to 1.

In Zero DCE, higher order curves are used to create more clear output images. Higher order curves are obtained by repeatedly applying the curves. The higher order curves are given by the equation.

$$LE_n(\mathbf{x}) = LE_{n-1}(\mathbf{x}) + \alpha_n LE_{n-1}(\mathbf{x})(1 - LE_{n-1}(\mathbf{x}))$$

The only driving force that moves this method toward correctness is a set of non reference loss functions . The model iterates over finding best fitting light enhancement curves and the aim of this iteration is to minimize the loss functions described below:

1. **Spatial Consistency Loss:** Spatial consistency refers to the contrast between adjacent pixels. This contrast is preserved by the curves being monotonous and the intensity of the resultant pixels is normalized to fall between [0,1]. The spatial consistency loss is defined by this equation:

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

where K = number of local regions, Y = average intensity value of the local region in the output, I = average intensity value of the local region in the input. Ω(i) = neighbouring regions of region denoted by i.

2. **Color Constancy Loss:** This loss is based on the theory that the average value off all pixels over all channels corresponds to grey and is used to rectify the color deviation in the input and the output. The color constancy loss is given by the following equation :

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

where J^p = average intensity of p channels in output, (p,q) = pair of channels.

3. **Exposure Control Loss :** This loss deals with the issue of over exposure or under exposure. We calculate the difference between the average value of local area and the value which we find to be well lit pixel value.The exposure control loss is given by the following equation :

$$L_{exp} = \frac{1}{M}\sum_{k=1}^{M}|Y_k - E|,$$

where M = number of regions, Y = average intensity value of a local region in output, E = The well exposed intensity value.

4. **Illumination Smoothness Loss :** This loss ensures that the transition between adjacent pixels is smooth and not abrupt.

$$L_{tv_\mathcal{A}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{c\in\xi}(|\nabla_x\mathcal{A}_n^c| + \nabla_y\mathcal{A}_n^c|)^2, \xi = \{R,G,B\}$$

where N = number of iterations, x = horizontal gradient, y = vertical gradient. The total loss is the sum of the above mentioned losses.

3.2. Zero-DCE++

Zero DCE ++ is a light weight version of zero-DCE, which works as efficiently but with very small number of hyperparameters (around 10K) and much less training time.

This version has eight times lesser hyper-parameters as compared to its other version that means the running time is almost half without compromising the quality of the results. According to the results mentioned in the paper, the inference speed of zero-DCE++ is very high(1000/11 FPS) on single CPU for image size upto 1200*900*3, whereas baseline DCE could only provide 500 FPS. On a i9 core 3.5 GHz CPU, DCE ++ could process an image of size 1200*900*3 within 1s whereas baseline DCE takes about 10s The FLOPS required by DCE for an image of same size were 84.99G while that of DCE++ were0.115G In some of the experiments we have done in our project, DCE++ has proven to be even better than the normal zero-DCE model.

3.3. Attention in Computer Vision

While looking into the possibilities for further improvement in the enhancement of images from current implementation baseline, one idea was tried which is to add attention module on top of the Zero-DCE++ model. In recent times, attention mechanisms are widely used especially in NLP domain. Attention is a method that tries to enhance the important parts while fading out the non-relevant information. It works similar to how human brain processes data. While processing images, human brain focuses on the specific parts and processes parts in a different priority.

Attention mechanisms were introduced into computer vision to mimic this aspect of the human visual system. So, attention mechanism can be considered as a dynamic weight adjustment process based on features of the input image.

Inspired by the multi-headed self attention mechanism from [19] for the NLP domain, the attention mechanism for computer vision domain was first proposed in the

paper Self Attention Generative Adversarial Networks (SAGAN), [27] that incorporated long range dependency capture for modelling image generation tasks.

There are two popular attention methods, which are used for computer vision tasks:

1. **Multi-Head Attention**: divides the features into heads and allows each attention module to focus only on a set of features. This gives greater power to encode multiple relationships and sub parts of images.
2. **Convolutional Block Attention Module (CBAM)** [25]: emphasizes meaningful features along the channel and spatial axes. The spatial attention module helps the model give more weight to the subject in each channel layer, whereas the channel attention module will help to identify important channels. The combination of both of these gives us a CBAM module as shown in Figure 2.

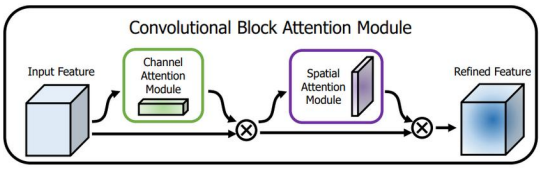

Figure 2. Layout of the CBAM Module

The attention mechanism used for the low-light image enhancement task is expected to improve the performance of the model especially, when there is a clear subject in the photograph. The proposed changes and corresponding results are mentioned in section 5.

4. Experimental Setup

4.1. Zero-DCE++

For the experimentation setup, Zero-DCE++ is taken as the baseline for low-light image enhancement task. The Zero-DCE++[1] model was trained and tested the available dataset[2] to generate the following results.

As we can see in the above figures, 4 is clearly enhanced compared to 3. The baseline model was trained using 2, P100 GPU and tested using an Intel Xeon CPU having 2.20 GHz clock speed.

4.2. Video Enhancement using Zero-DCE++

As we have already mentioned, Zero-DCE is lightweight and it's inference time is quite less, so we want to explore the quality and the processing time needed when it's used for video enhancement. Here, Zero-DCE++ was used, which is an accelerated and even lighter version of Zero-DCE.

Figure 3. Low-light input image

Figure 4. Enhanced output image using Zero-DCE++

For this experiment, a low-light video was taken as input and every frame of it was enhanced by passing those through the Zero-DCE++ model and finally the frames were combined to get the enhanced video. The experiment is being performed in an Intel Xeon CPU having 2.20 GHz clock speed. The code[3], one original video[4] and corresponding enhanced video[5] are present in the following locations.

4.3. Real-Time Video Enhancement

Next attempt was to carry out video enhancement on real-time streams from camera. Live laptop web camera feed is used as input and trained Zero-DCE model is used for generating the enhanced frames. After training the model, the model will get stored in the snapshot folder. Real-time video enhancement module captures the real-time video frames from laptop camera, runs Zero-DCE enhancement on the frames of captured video and displays enhanced video along with the original video. Here are the

[1] https://github.com/Tamal-Mondal/Low_Light_Image_Enhancement_2

[2] https://tinyurl.com/3edempds

[3] https://tinyurl.com/4axm9m3e

[4] https://tinyurl.com/2p9xrhmp

[5] https://tinyurl.com/3dczyb7j

screenshots of the original video and enhanced video.

Figure 5. Real-time video enhancement(left-original video , right-enhanced video)

The code[6], one original baseline real-time video[7] and corresponding enhanced video[8] are present in the following locations.
This module was tested on AMD machine with clock speed of 2.60 GHz and no GPUs used. System was able to produce output in speed of 10FPS. Zero-DCE++ model with 100 epoch model is used to enhance the real time video.

5. Proposed approach: Zero-DCE++ with CBAM Attention

This section describes the details of various improvements that were attempted over Zero-DCE++. Attempts were to improve the visual quality of the enhanced image without increasing the inference time, as this model is meant to be used in the real-time applications. We tried to incorporate simple approaches without complicating the model's internal structure. Attention on existing model is one of the approach, that we tried. Various models were trained by changing various hyper parameters of the model. Detailed descriptions of those attempted steps are described below.

We have chosen to experiment with Convolutional Block Attention Module(CBAM) [25] that combines both channel and spatial attention. Standard implementation of CBAM [17] is used for our experiment and have used both channel and spatial attention together to improve the enhanced images as suggested by the authors in the original CBAM paper. This CBAM attention module is added over each convolution layer's activation maps to generate each corresponding feature maps, that focuses on relevant and dependent pixels (area of focus) of the image over all the channels for the low-light image enhancement task.

The figure 6 shows the architecture of the model after combining Zero-DCE++ with CBAM. One CBAM layer was added on top of each of the 7 layers of original Zero-DCE++ model. We have used two different types of CBAM layers, first one with reduction ratio 4 that is applied on first 6 CNN layers and second one with reduction ratio 2 (as there is only 3 channels in final CNN layer output) which is applied on the final CNN layer. We have used only average and maxpool together(used in channel attention part of CBAM) for all CBAM layers as suggested in original CBAM paper. We have got best output when we haven't applied batch norm and have applied bias in CNN layer of CBAM(used in special attention part of CBAM).

We have conducted a variety of experiments with CBAM to understand how different hyper-parameters effect model's performance and it is all mentioned in the Ablation study section 7. The results and it's analysis after the modifications is mentioned in section 6 in detail.

[6] https://tinyurl.com/2p8fzd47
[7] https://tinyurl.com/2p8zzptw
[8] https://tinyurl.com/59x7mc39

Model	PSNR↑	SSIM↑	MAE↓
Zero-DCE++ (baseline)	11.52	0.062	66.19
CBAM in first 6 layers (*A*)	13.22	0.062	55.19
CBAM(*A*) with bias and no batch_norm (*B*)	11.98	0.062	63.19
CBAM(*B*) with 4 pooling types	**13.45**	0.062	**42.88**
CBAM in all 7 layers with bias and no batch_norm (*C*)	9.10	0.062	87.39
CBAM(*C*) with reduced reduction_rate (*D*)	8.66	0.062	92.02
CBAM(*D*) with wd=0.001 and lr=0.001	6.04	0.055	122.54
CBAM(*D*) in first 4 layers	10.38	0.055	75.42

Table 1. Evaluation metric (Average PSNR, Average SSIM and Average MAE) results for different frameworks upon baseline. *The letters at the end of model description is there to give label to that particular framework to be used in subsequent models.*

6. Results and Analysis

6.1. Quantitative

6.1.1 Image Quality Assessment Metrics

The image quality assessment metrics PSNR, SSIM, and MAE metrics are used to quantitatively compare the performance of different methods. A higher SSIM value indicates a result is closer to the ground truth in terms of structural properties. A higher PSNR (lower MAE) value indicates a result is closer to the ground truth in terms of pixel level image content. The table 1 gives the details of the metrics on different approaches.

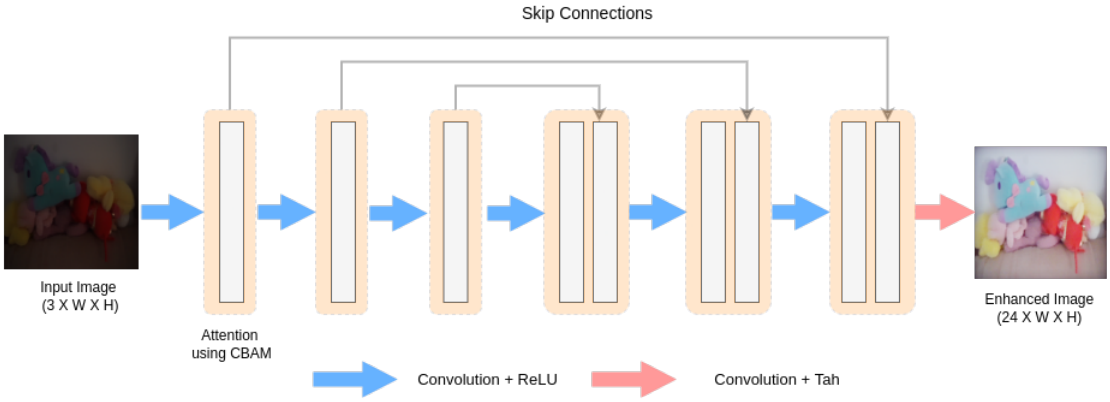

Figure 6. CBAM Attention over DCE-Net Architecture

6.1.2 Training Loss and Testing Time Comparisons

This section describes details on the training and test loss trends of various approaches.

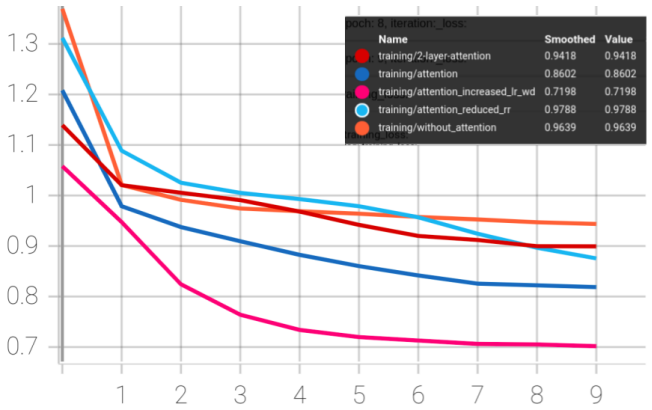

Figure 7. Training Loss over different experiments (Smoothing = 0)
X-axis: Epoch, Y-axis: Overall Training Loss

Looking at the training loss curve as shown in figure 7, the training loss for the baseline implementation (labelled as *training/without_attention*), which is the Zero-DCE++ is 0.9437 for 10 epoch where as the training loss for the attention model (CBAM attention over DCE-Net) is 0.8602 over the same number of epochs. Upon a number of additive experiments with attention model we have the best performing model (labelled as *training/attention_reduced_rr*)having reduction ratio changed from 16 originally to 4 for the first size layers and 2 for the last/output layer, selected from experimental analysis. It is clearly seen that the best performing model is having a significantly lesser training loss with better/early convergence. This result infers that the model is learning well and consistently performing well on the loss functions defined.

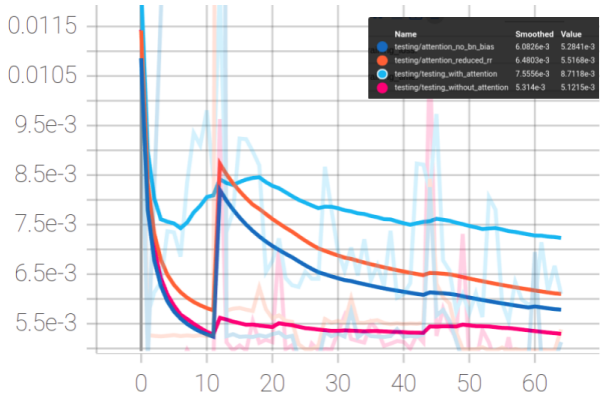

Figure 8. Testing time for different images over different models (Smoothing = 0.99)
X-axis: Images, Y-axis: Overall Testing Time

Similarly, looking at the testing time over different model configurations, the baseline model (with no any attention) is performing the fastest. With addition of CBAM attention modules, there has been increase in number of learnable parameters along with the number of matrix calculations that resulted in the increase in testing time.

The complete set of experiments with quantative results is published here. [9]

6.2. Qualitative

The figure 9 shows the qualitative comparison between different frameworks discussed in this paper on the same image. It provides a more practical view on the performance of the model which is difficult to capture in terms of quantitative metrics. Another figure (Figure 10) shows the comparison between two different improved frameworks along with the baseline model results on an image that is captured locally from the camera of a smartphone.

[9] https://tinyurl.com/36z5keu2

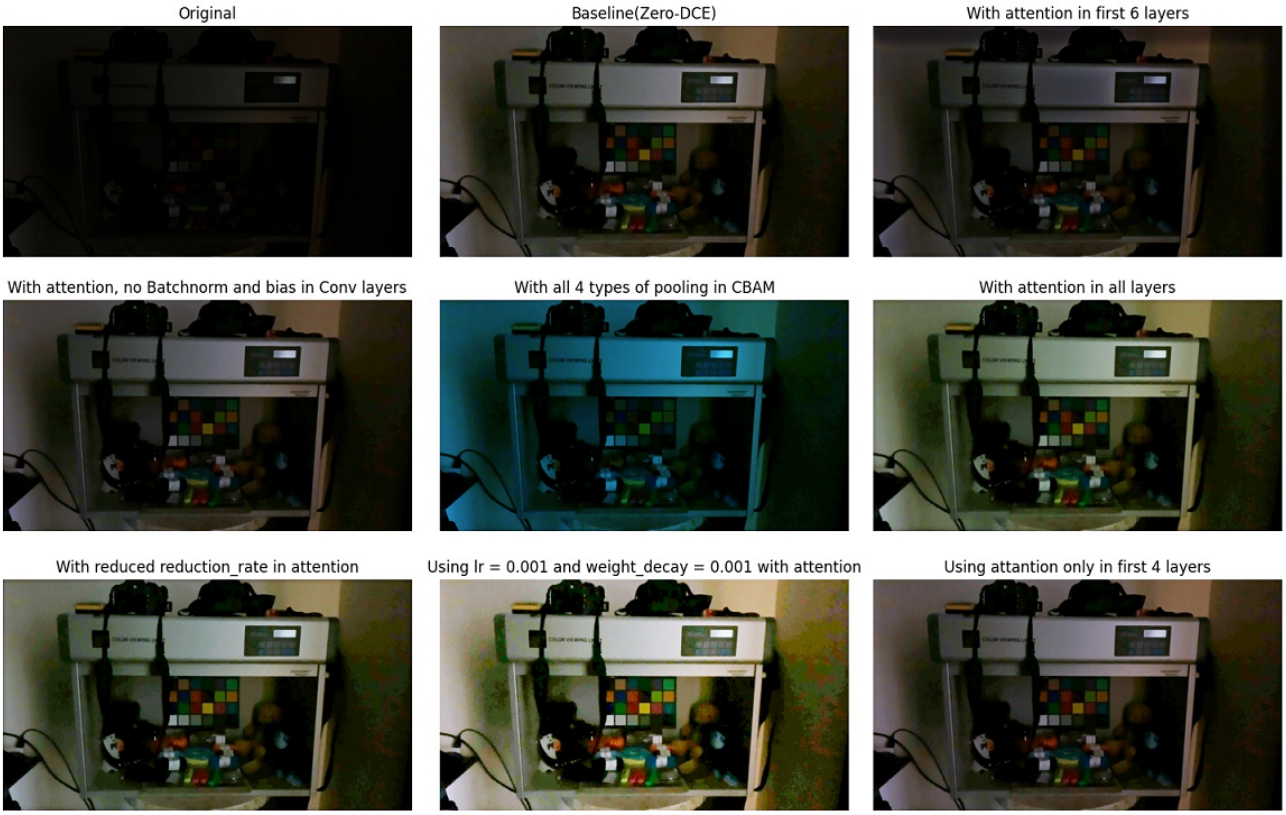

Figure 9. Qualitative comparison between different frameworks

Figure 10. Qualitative comparison between different frameworks for a local camera captured image

7. Ablation Study

Figure 11 shows the effect of changing number of attention layers on enhanced output image. We have experimented by applying attention in first 4, 6 and 7(all layers including output layer) respectively. We can see when we applied attention to all the layers, we have got the best results followed by attention in 6 layers and 4 layers respectively.

7.1. Effect of applying Batch norm and bias in CBAM

We know that batch norm can speed up the training but during that it can also break the relations of neighboring pixels which can effect the model output. Similarly adding bias in the convolution layers can also help to better train the model. We have observed that during our experiments and figure 12 depicts that. We have got our best results by not applying batch norm and adding bias in CNN layer of CBAM.

7.2. Effect of reduction rate of CBAM

In the original CBAM paper, 16 was mentioned as a good reduction rate for channel attention layer in CBAM but we believe it depends on the problem. With increase in the reduction rate, number of neurons decreases in the linear layer of channel attention. We have got best results when we used reduction rate 4 for first 6 attention layers and 2 for last attention layer(as number of channels are just 3). Figure 13 shows the effect on output image.

Figure 11. Effect of attention on different number of layers of Zero-DCE++

Figure 12. Effect of adding bias and removing batch_norm in Attention enhanced Zero-DCE++ (*left: original attention framework*)

Figure 13. Effect of reducing reduction rate in Attention enhanced Zero-DCE++ (*left: original attention framework*)

7.3. Effect of different types of pooling of CBAM

As also suggested in the original CBAM paper, when we used average and max pooling together in channel attention layer of CBAM, we have got the best results. With this 2 we also tried adding power average pooling(LP-Pool) and logarithmic summed exponential(LSE) but that deteriorates the output image as we can see in figure 14.

7.4. Effect of learning rate and weight decay

In original Zero-DCE model, authors used learning rate and weight decay(for optimizer) as 0.0001 which even though quite small, has given pretty good results to us with just 10 epochs. When we increased the value to 0.001, we observed that loss decreases more as expected but output was little noisy or over-exposed in some cases as we can see in the Figure 15.

8. Criticism

Here are some of the downside of our approach that we would like to highlight and we will try to address these in future.

- **Increase in training time**: As we have added attention layers after each convolution layer, it definitely increases the training time. Using 2, P100 GPU we have experimented and have seen that after modifications, the model with attention takes roughly 5 times more training time with respect to baseline Zero-DCE++ model.

  We have noticed that with batch norm in the attention layer, increase in training time is not this much but that also effects the results as we have mentioned previously.

- **Saturation**: There are cases where we have observed that attention model decreases the color saturation of the output image, especially in the cases where the texture of the image is light as we can see in figure 16.

- **Discrepancies between quantitative and qualitative results**: More the PSNR and SSIM is better the output image and similarly less the MAE is better for output. But we think these are not that good matrices to evaluate the outputs of low-light image enhancement task where quality is subjective. As we can see in the quantitative results, our best model(7 layers of attention, no batch norm and reduced reduction rate) has lower PSNR(8.66 vs 11.52) and higher MAE(92.03 vs 66.19) than Zero-DCE++ but the output images look better to the human eye.

Figure 14. Effect of adding all 4 types of pooling in Attention enhanced Zero-DCE++ (*left: original attention framework*)

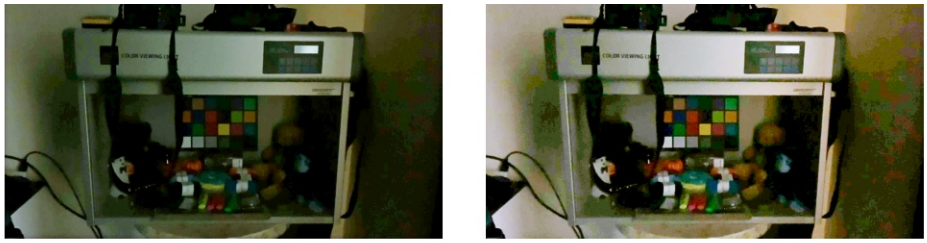

Figure 15. Effect of increasing learning_rate and weight_decay in Attention enhanced Zero-DCE++ (*left: original attention framework*)

Figure 16. Difference in color saturation in the Zero-DCE++ baseline and the Attention enhanced model.

9. Conclusions

Present work on low light image enhancement using light enhancement curves opened up a new way of improving the illumination of low light images using deep learning approach. As part this work, ZeroDCE++ based approach was taken as the baseline and various approaches were attempted out to improve the image quality without degrading the performance of the baseline model. Attention based approaches on baseline ZeroDCE++ model improved the quality of the enhanced image. There are still improvements required in the extreme low light image enhancement, in terms of the noise corrections, simplifying the model complexity in order to achieve good performance in the real-time applications. Since, various vision based projects requires more efficient pre-processing pipeline, improved low light image enhancement approaches are welcomed always.

The existing work using zero-DCE++ may be explored to achieve better performance in terms of visual quality, noise correction, inference time, training time and other artifacts.

10. Credit Statement

- **Problem Identification**: Collaborative Effort
- **Literature Review**: Collaborative Effort
- **Proposed Approach**:
  - Implementation of Attention mechanism: Tamal Mondal , Kamal Shrestha
  - Incorporation of CBAM based model: Tamal Mondal
  - Video Enhancement: Tamal Mondal
  - Realtime Video Enhancement : Jayamohan.C.B
- **Fine Tuning hyper parameters**:
  - Effect of multiple attention layers: Kamal Shrestha, Tamal Mondal
  - Effect of Pooling: Aman Agrawal, Tamal Mondal
  - Effect of Batch norm and bias: Kamal Shrestha, Praveen Vishwakarma
  - Effect of learning rate and Weight Decay: Tamal Mondal , Jayamohan.C.B
- **Tensorboard Evaluations and ML-Flow checkpoints**: Kamal Shrestha

- **Model Testing**: Aman Agrawal, Kamal Shrestha, Praveen Vishwakarma
- **Configuration Management**: Tamal Mondal , Kamal Shrestha
- **Documentation**: Collaborative Effort

11. Acknowledgements

This work was carried out as part of Deep Learning course(AI5100) offered at IIT Hyderabad. We would like to thank Dr. Sumohana S. Channappayya(Professor, Department of Electrical Engineering, IIT Hyderabad) who has guided us throughout the project and have given valuable advises whenever necessary.

References

[1] Partha Pratim Banik, Rappy Saha, and Ki-Doo Kim. Contrast enhancement of low-light image using histogram equalization and illumination adjustment. In *2018 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–4, 2018. 2

[2] Zilong Chen, Yaling Liang, and Minghui Du. Attention based broadly self-guided network for low light image enhancement, 2021. 3

[3] Chi-Mao Fan, Tsung-Jung Liu, and Kuan-Hsien Liu. Half wavelet attention on m-net+ for low-light image enhancement, 2022. 3

[4] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2782–2790, 2016. 2

[5] 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 (CVPR)*, June 2020. 1

[6] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017. 3

[7] Haidi Ibrahim and Nicholas Sia Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(4):1752–1758, 2007. 2

[8] 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. 1, 3

[9] Mohit Lamba and Kaushik Mitra. Restoring extremely dark images in real time. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3486–3496, 2021. 3

[10] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977. 2

[11] C. Li, C. Guo, L. Han, J. Jiang, M. Cheng, J. Gu, and C. Loy. Low-light image and video enhancement using deep learning: A survey. *IEEE Transactions on Pattern Analysis Machine Intelligence*, (01):1–1, nov 5555. 2, 3

[12] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 4

[13] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. 2

[14] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *CoRR*, abs/1511.03995, 2015. 3

[15] Feifan Lv, Yu Li, and Feng Lu. Attention guided low-light image enhancement with a large scale low-light simulation dataset, 2019. 3

[16] Raghav Mehta and Jayanthi Sivaswamy. M-net: A convolutional neural network for deep brain structure segmentation. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 437–440, 2017. 3

[17] Diganta Misra. Attention mechanisms in computer vision: Cbam, 2AD. 7

[18] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing, 2022. 3

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[20] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6849–6857, 2019. 1, 3

[21] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, 22(9):3538–3548, 2013. 3

[22] Wencheng Wang, Xiaojin Wu, Xiaohui Yuan, and Zairui Gao. An experiment-based review of low-light image enhancement methods. *IEEE Access*, 8:87884–87917, 2020. 2

[23] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex C. Kot. Low-light image enhancement with normalizing flow, 2021. 3

[24] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 1, 2, 3

[25] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 6, 7

[26] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *European Conference on Computer Vision*, pages 492–511. Springer, 2020. 4

[27] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 6

[28] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1