Low Light Image Enhancement using MIRNet

A Mini Project in Image Processing(EC386) report
Submitted in Partial Fulfillment of the Requirements for the Degree of

BACHELOR OF TECHNOLOGY

in

Electronics and Communication Engineering

by

Reuben Silveira
(211EC139)

Chandan Kumar
(211EC108)

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERINGNATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
SURATHKAL, MANGALORE - 575025
November 2023

CERTIFICATE

This is to certify that the Mini Project Report entitled Low Light Image Enhancement using MIRNet submitted by Reuben Silveira (Register No: 211EC139) and Chandan Kumar (Reg no: 211EC108) as a record of the Mini Project in Image Processing (EC386) is accepted in partial fulfilment of the requirements for the award ofBachelor of Technology Degree in the Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal, Mangaluru.

Project Guide: Dr. Sumam David

Course Instructor: Dr. Shyam Lal

ABSTRACT

Image restoration which aims to recover high quality content from degraded images finds applications in various fields such as surveillance, computational photography, medical imaging, and remote sensing. Convolutional Neural Networks (CNNs) have significantly outperformed traditional methods in image restoration tasks. Current CNN-based approaches operate either on full-resolution yielding spatially precise but less contextually robust results, or on progressively low-resolution representations, producing contextually reliable but spatially less accurate outputs. This project implements a novel architecture that addresses these limitations by maintaining spatial precision throughout the network while utilizing strong contextual information from low-resolution representations.

The core of our approach is a multi-scale residual block that includes several crucial elements:
(a) parallel multi-resolution convolution streams for extracting features at different scales
(b) information exchange among these multi-resolution streams
(c) spatial and channel attention mechanisms to capture contextual information
(d) attention-based multi-scale feature aggregation.

In summary our method named MIRNet, learns an enriched set of features by combining contextual information from multiple scales while simultaneously preserving high resolution spatial details. Simulation results on test dataset and evaluation metrics reveal that the model achieves remarkable results for image processing tasks in the case of low light image enhancement.

Contents

List of Tables

List of Figures

1 Introduction

The proliferation of cameras in various devices has led to an exponential growth in image content. However, during image acquisition degradation often occurs due to physical camera limitations or inappropriate lighting conditions. For instance smartphone cameras with narrow apertures and small sensors frequently produce noisy and low-contrast images. Additionally images captured in unsuitable lighting conditions may be either too dark or too bright. The challenge of recovering the original clean image from such corrupted measurements is studied in the field of image restoration.

Recent advancements in deep learning models have significantly improved image restoration and enhancement by learning strong priors from large-scale datasets. Existing CNNs typically follow one of two architecture designs: encoder-decoder or high-resolution (single-scale) feature processing. Encoder-decoder models progressively map input to low-resolution representations and then apply a reverse mapping to the original resolution [18,19]. While these approaches capture a broad context, they often lose fine spatial details. On the other hand high-resolution (single-scale) networks preserve spatially accurate details but struggle to encode contextual information due to a limited receptive field [20,21].

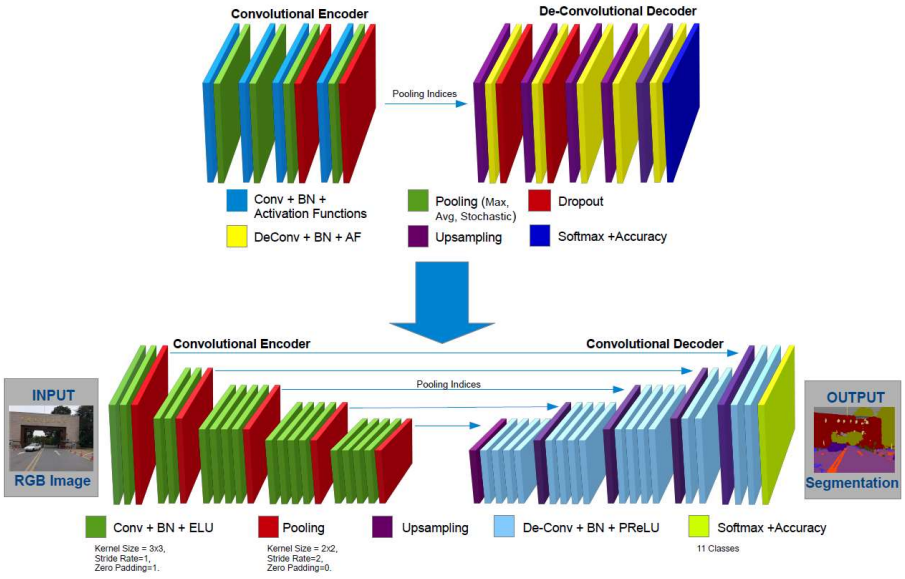


Figure 1: Encoder - Decoder Type CNN Network

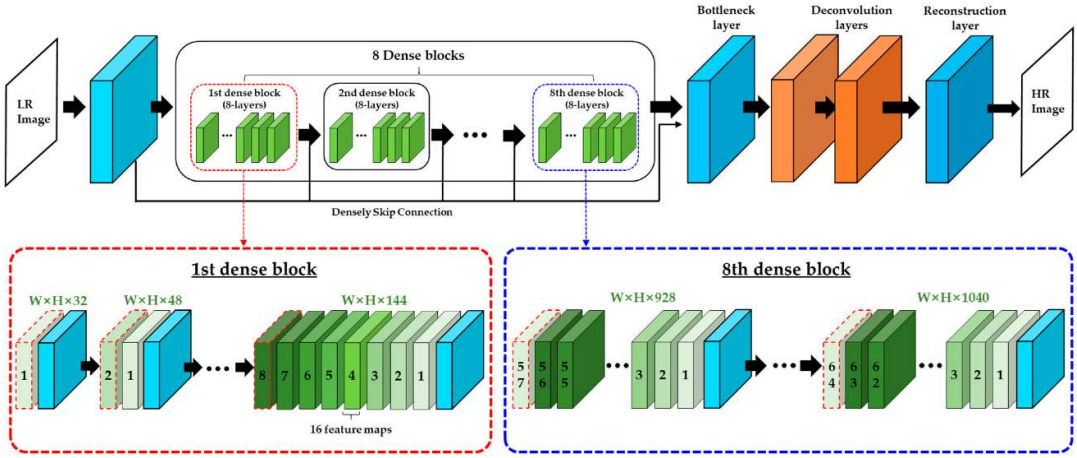


Figure 2: High Resolution (Single Scale) Network

Image restoration requires position sensitive procedures to maintain pixel-to-pixel correspondence between input and output images. It is crucial to remove only undesired degraded content while preserving desired fine spatial details, such as true edges and texture. To achieve this we implement a novel multi-scale approach that maintains original high-resolution features throughout the network hierarchy to minimize the loss of precise spatial details. Simultaneously our model encodes multi-scale context using parallel convolution streams that process features at lower spatial resolutions. The multi-resolution parallel branches operate complementarily to the main high-resolution branch providing more precise and contextually enriched feature representations.

Our method differs from existing multi-scale image processing approaches in the way it aggregates contextual information. While existing methods process each scale in isolation and exchange information only in a top-down manner our approach progressively fuses information across all scales at each resolution level allowing both top-down and bottom-up information exchange [22,23]. The selective kernel fusion mechanism facilitates lateral knowledge exchange between fine-to-coarse and coarse-to-fine scales on each stream.

Unlike methods that use simple concatenation or averaging, our fusion approach dynamically selects useful kernels using a self-attention approach. The proposed fusion block combines features with varying receptive fields while preserving their distinctive complementary characteristics.

The main features of this implementation include:

- Introduction of an innovative feature extraction model that acquires a complementary set of features across multiple spatial scales while simultaneously preserving the original high-resolution features to retain precise spatial details.

- Implementation of a systematically repeated mechanism for information exchange where features from multi-resolution branches are progressively merged to enhance representation learning.

- Adoption of a fresh approach to fuse multi-scale features using a selective kernel network that dynamically combines variable receptive fields ensuring the faithful preservation of original feature information at each spatial resolution.

- Introduction of a recursive residual design that systematically dissects the input signal to streamline the overall learning process enabling the construction of highly deep network

2 Literature Review

Due to rapidly growing image content there's an urgent need to develop effective image restoration and enhancement algorithms. In this project we implement a new method capable of performing image enhancement. This approach processes features at the original resolution in order to preserve spatial details while effectively fuses contextual information from multiple parallel branches.

Traditional Spatial Domain Techniques: Early image enhancement methods focused on spatial domain techniques. Landini and Rand's work, "Image Enhancement in the Spatial Domain" (1975) [2], laid the groundwork for manipulating pixel intensities and spatial features.

Retinex Theory for Image Decomposition: McCann's introduction of Retinex theory in "Lightness and Retinex Theory" (1971) [3] became pivotal for separating illumination and reflectance components enhancing image decomposition strategies.

Deep Learning Dominance: Chen and Zhang's "Learning a Deep Convolutional Network for Image Super-Resolution" (2015) [4] exemplifies the transformative role of deep learning in capturing complex features directly from data.

Exposure Fusion for Low Light: Exposure fusion techniques, as explored by Mertens et al. in "Exposure Fusion" (2007) [5], have been adapted for low light scenarios, combining multiple images at different exposures for enhanced results.

Noise Reduction Strategies: Addressing noise challenges in low light images, Wiener's principles from "Extrapolation, Interpolation, and Smoothing of Stationary Time Series" (1964) [6] have inspired effective noise reduction methods.

For image enhancement, histogram equalization is the most commonly used approach. However, it frequently produces under- or over-enhanced images. Motivated by the Retinex theory [7], several enhancement algorithms mimicking human vision have been proposed in the literature [8,9,10]. Recently CNNs have been successfully applied to general, as well as low-light, image enhancement problems [11]. Notable works employ Retinex-inspired networks [12,13], encoder-decoder networks [14,15], and GANs [16,17].

These diverse approaches together contribute to the comprehensive landscape of image enhancement showcasing a transition from traditional spatial domain techniques to the adoption of sophisticated deep learning methods.

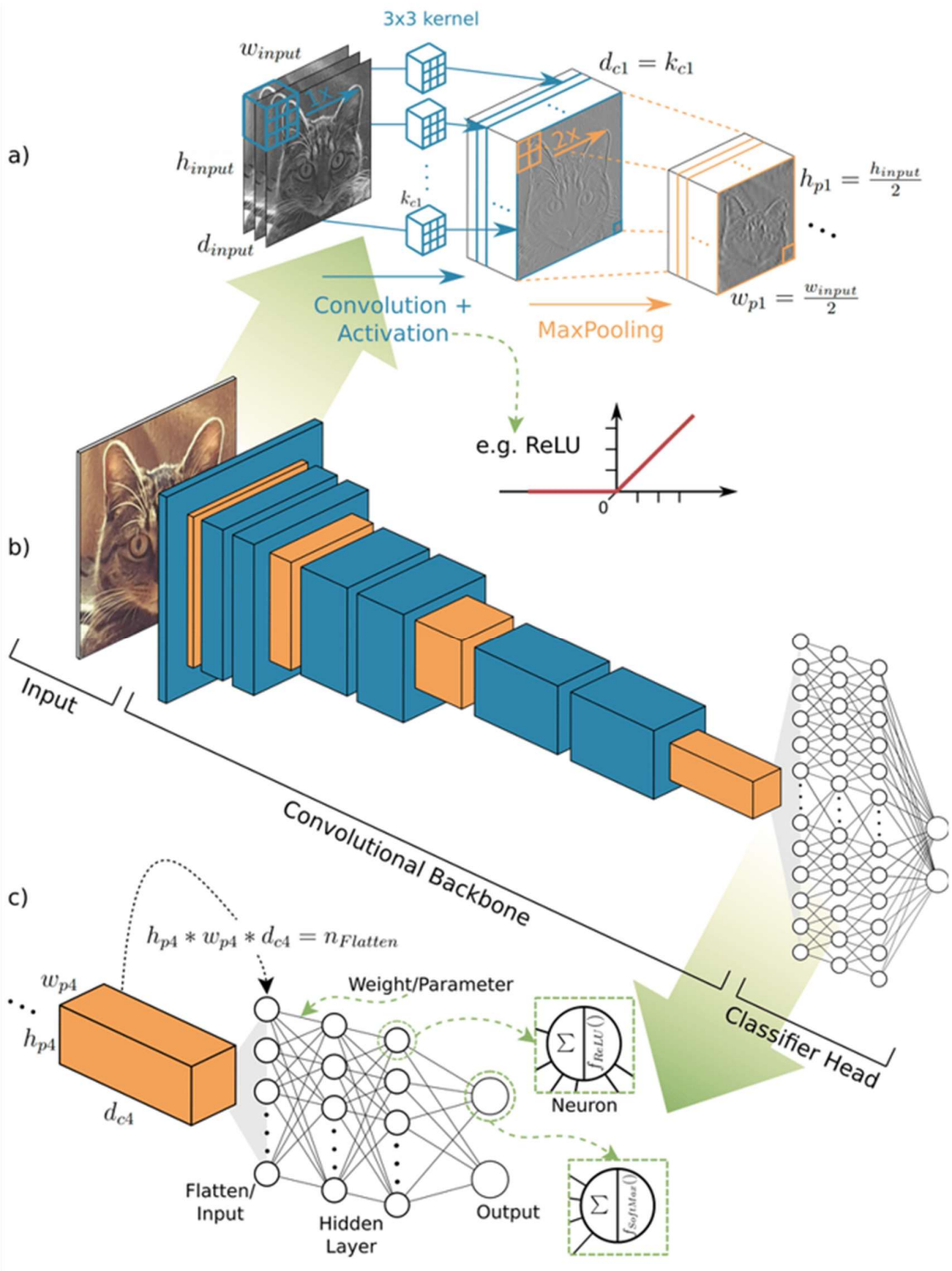


Figure 3: General CNN Architecture

3 Methodology

3.1 Data Preparation and Preprocessing

The LoL (LOw-Light) dataset is used for training and evaluating the MIRNet model. The dataset which is readily available in a public google drive link is downloaded and then unzipped. The dataset contains two sets of images: low-light images and their corresponding enhanced (brightened) versions. Out of 500 pairs of low and enhanced images each 400x600 in size, 300 are used as training data, 185 for validation purposes and remaining 15 as test data. The image pairs are then organised together. Training data is used to update the model's parameters during training while validation data helps assess the model's performance and prevent overfitting

Data preprocessing involves creating random crops of the images which are 128x128 in size and converting them to a consistent format. The images are read and decoded from their file paths using TensorFlow functions so that they are in a suitable format for further processing

3.2 Dataset Creation

We use TensorFlow datasets to feed the model during training and validation. In order to create these datasets we first define functions that read and load images from their file paths. The image file path is given as input to the function which reads the image file, then decodes it (it's assumed to be in PNG format with three color channels RGB) and also normalizes the pixel values to be between 0 and 1

Another function is used for creating random sections (crops) from the images, this is a technique used for data augmentation. It randomly selects a portion of the low-light image and its corresponding enhanced image both of size 128x128 pixels. This cropping is performed to add diversity to the training data and helps the model

generalize better.

We use the load_data function, the read_image and random_crop functions to load and preprocess the low-light and enhanced images. It reads both images, applies random cropping and returns the resulting cropped images as pairs.

The images are processed in batches to improve training efficiency. The datasets

are batched with a batch size of 4 which means that each training step will use four image pairs.

3.3 Model Architecture Design

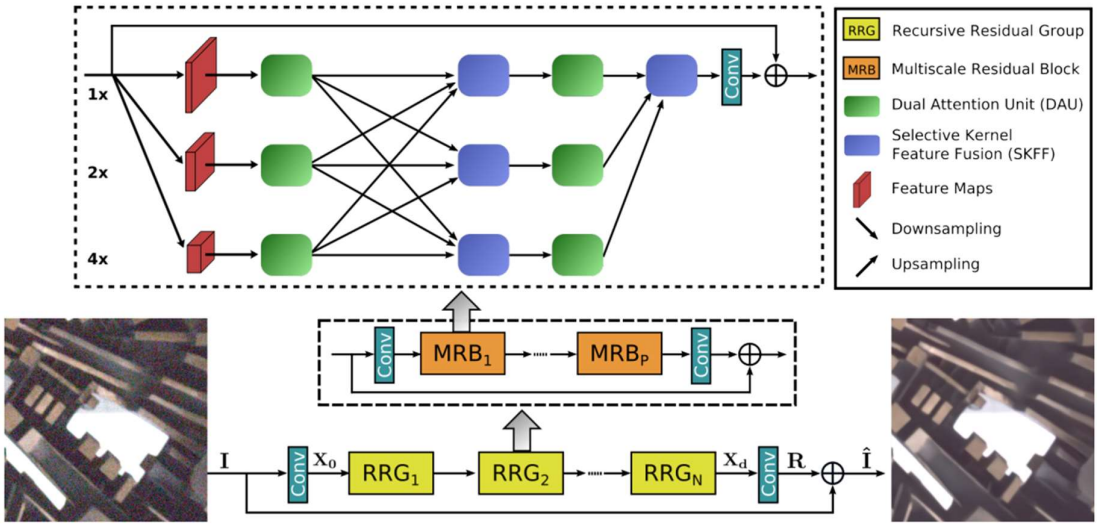


Figure 4: Framework of The Proposed MIRNet Network

The main features of the MIRNet model are:

- A feature extraction model that computes a complementary set of features across multiple spatial scales, while maintaining the original high-resolution features to preserve precise spatial details.
- A regularly repeated mechanism for information exchange, where the features across multi-resolution branches are progressively fused together for improved representation learning.
- A new approach to fuse multi-scale features using a selective kernel network that dynamically combines variable receptive fields and faithfully preserves the original feature information at each spatial resolution.
- A recursive residual design that progressively breaks down the input signal in order to simplify the overall learning process, and allows the construction of very deep networks.

3.3.1 Selective Kernel Feature Fusion

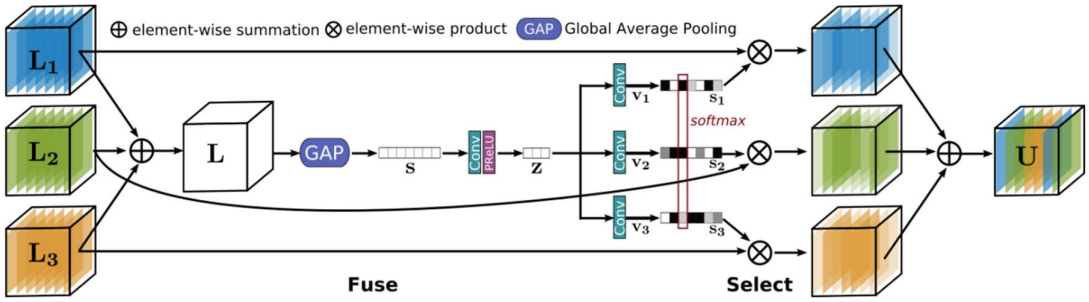


Figure 5: Schematic of Selective Kernel Feature Fusion (SKFF)

The Selective Kernel Feature Fusion or SKFF module performs dynamic adjustment of receptive fields via two operations: Fuse and Select. The Fuse operator generates global feature descriptors by combining the information from multi-resolution streams. The Select operator uses these descriptors to recalibrate the feature maps (of different streams) followed by their aggregation.

Fuse: The SKFF receives inputs from three parallel convolution streams carrying different scales of information. We first combine these multi-scale features using an element-wise sum, on which we apply Global Average Pooling (GAP) across the spatial dimension. Next, we apply a channel- downscaling convolution layer to generate a compact feature representation which passes through three parallel channel-upscaling convolution layers (one for each resolution stream) and provides us with three feature descriptors.

Select: This operator applies the softmax function to the feature descriptors to obtain the corresponding activations that are used to adaptively recalibrate multi-scale feature maps. The aggregated features are defined as the sum of product of the corresponding multi-scale feature and the feature descriptor.

3.3.2 Dual Attention Unit

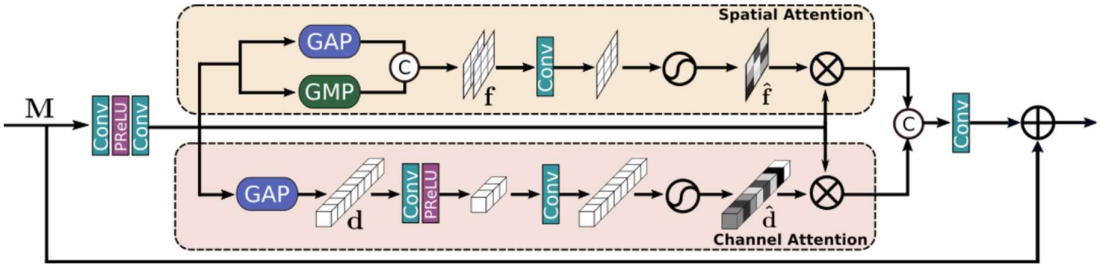


Figure 6: Dual Attention Unit Schematic

The Dual Attention Unit or DAU is used to extract features in the convolutional

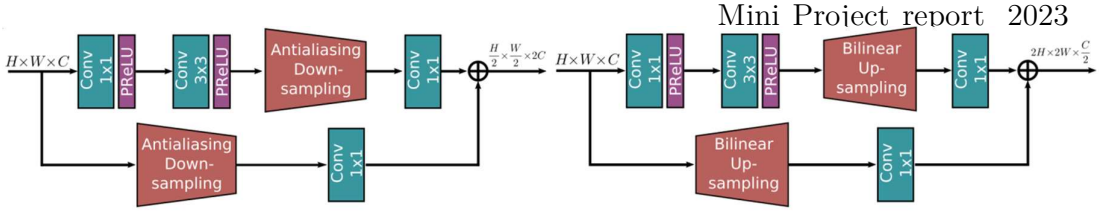
streams. While the SKFF block fuses information across multi-resolution branches, we also need a mechanism to share information within a feature tensor, both along the spatial and the channel dimensions which is done by the DAU block. The DAU suppresses less useful features and only allows more informative ones to pass further. This feature recalibration is achieved by using **Channel Attention** and **Spatial Attention** mechanisms.

The **Channel Attention** branch exploits the inter-channel relationships of the convolutional feature maps by applying squeeze and excitation operations. Given a feature map, the squeeze operation applies Global Average Pooling across spatial dimensions to encode global context, thus yielding a feature descriptor. The excitation operator passes this feature descriptor through two convolutional layers followed by the sigmoid gating and generates activations. Finally, the output of Channel Attention branch is obtained by rescaling the input feature map with the output activations.

The **Spatial Attention** branch is designed to exploit the inter-spatial dependencies of convolutional features. The goal of Spatial Attention is to generate a spatial attention map and use it to recalibrate the incoming features. To generate the spatial attention map, the Spatial Attention branch first independently applies Global Average Pooling and Max Pooling operations on input features along the channel dimensions and concatenates the outputs to form a resultant feature map which is then passed through a convolution and sigmoid activation to obtain the spatial attention map. This spatial attention map is then used to rescale the input feature map.

3.3.3 Multi-Scale Residual Block

The Multi-Scale Residual Block is capable of generating a spatially-precise output by maintaining high-resolution representations, while receiving rich contextual information from low-resolutions. The MRB consists of multiple (three in this project) fully-convolutional streams connected in parallel. It allows information exchange across parallel streams in order to consolidate the high-resolution features with the help of low-resolution features, and vice versa. The MIRNet employs a recursive residual design (with skip connections) to ease the flow of information during the learning process. In order to maintain the residual nature of our architecture, residual resizing modules are used to perform downsampling and upsampling operations that are used in the Multi-scale Residual Block.



(a) Downsampling module (b) Upsampling module

Figure7: Residual Resizing Modules to Perform Downsampling and Upsampling

3.4 Loss Function, Optimizer and Pipeline

Let $\mathbf{I}$ be an Image of dimensions $\mathbb{R}^{HxWx3}$. The network first applies a convolutional layer to extract low-level features $\mathbf{X_0} \in \mathbb{R}^{HxWxC}$.

Next the feature maps $\mathbf{X_0}$ to pass through N number of recursive residual groups (RRGs) which yield deep features $\mathbf{X_d} \in \mathbb{R}^{HxWxC}$. RRG contains several multi-scale residual blocks.

In the next step we apply one more convolutional layer to deep features $\mathbf{X_d}$ to obtain a residual image $\mathbf{R} \in \mathbb{R}^{HxWx3}$.

The restored image is obtained as follows: $\hat{\mathbf{I}} = \mathbf{I} + \mathbf{R}$.

We use Charbonnier loss to optimize our proposed network.

$$\mathcal{L}(\hat{\mathbf{I}}, \mathbf{I}^*) = \sqrt{\left\|\hat{\mathbf{I}} - \mathbf{I}^*\right\|^2 + \varepsilon^2} \tag{1}$$

Where $\mathbf{I^*}$ denotes the ground-truth image

ε is a constant which we empirically set to 10^{-3} for all the experiment.

3.5 Model Compilation

The loss function measures the difference between the predicted output and the ground truth (target) for a given input during training. For the MIRNet model a custom loss function called charbonnier loss is defined. This function calculates the loss by computing the root mean square error with a small constant added.

An optimizer is used to adjust the model's learnable parameters (weights and biases) during training and minimise

the loss function. Here, we use the "Adam" optimizer with a specified learning rate of (1e-4).

The Peak Signal-to-Noise Ratio (PSNR) is used as the evaluation metric. PSNR measures the quality of the model's output by comparing it to the ground truth. Higher PSNR values indicate better image quality and less distortion.

$$PSNR = 10\log\left(\frac{\text{MAX}^2}{\text{MSE}}\right)$$

- MAX represents the maximum possible pixel value of the image (for example, for an 8-bit grayscale image, MAX = 255).
- MSE is the Mean Squared Error between the original and distorted images.

3.6 Model Training

The training was using Keras library of tensorflow. We used adam optimizer for training with a learning rate of 1e-4. Learning rate plateau was applied to reducethe learning rate according to val loss improvement. Training deep learning models is an iterative process and each iteration is called an "epoch." The number of epochs specifies how many times the entire training dataset is processed by the model. In this project the number of epochs is set to 50.

Learning rate reduction callbacks are a mechanism to adapt the learning rate during training. we are using a learning rate reduction callback to monitor the validation PSNR (Peak Signal-to-Noise Ratio). If the validation PSNR does not improve for a certain number of epochs then the learning rate is reduced by a factor specified in the callback.

During each epoch the MIRNet model processes mini-batches of low-light images and their targets from the training dataset. It computes the loss using the charbonnier loss function and backpropagates the gradients through the network. The optimizer (Adam) then updates the model's learnable parameters to minimize the loss.

The model's performance is monitored throughout training. The training loss, validation loss, and PSNR (Peak Signal-to-Noise Ratio) are tracked over the epochs. The training loss should decrease indicating that the model is learning to restore low-light images.

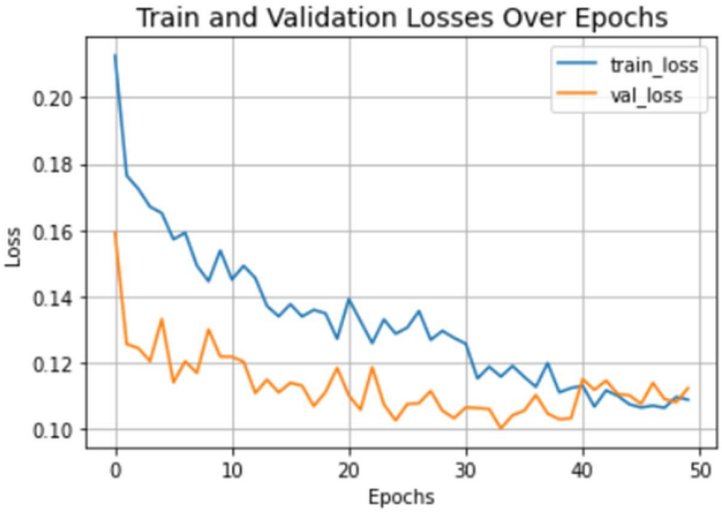


Figure 8: Train and Validation Losses Over Epochs

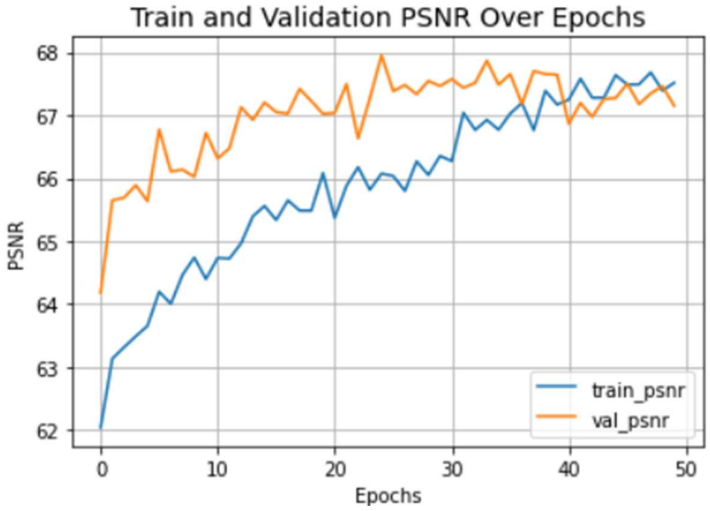


Figure 9: Train and Validation PSNR Over Epochs

4 Discussions

4.1 LoL Dataset

The LoL test dataset images were given as input to the model after the completion of training phase and the outputs are compared with the ground-truth image and with another enhanced version of the low light image obtained via the Python Imaging Library(PIL) using its ImageOop.autocontast(). Out of the 15 test images here we display the results of 5 of these images.

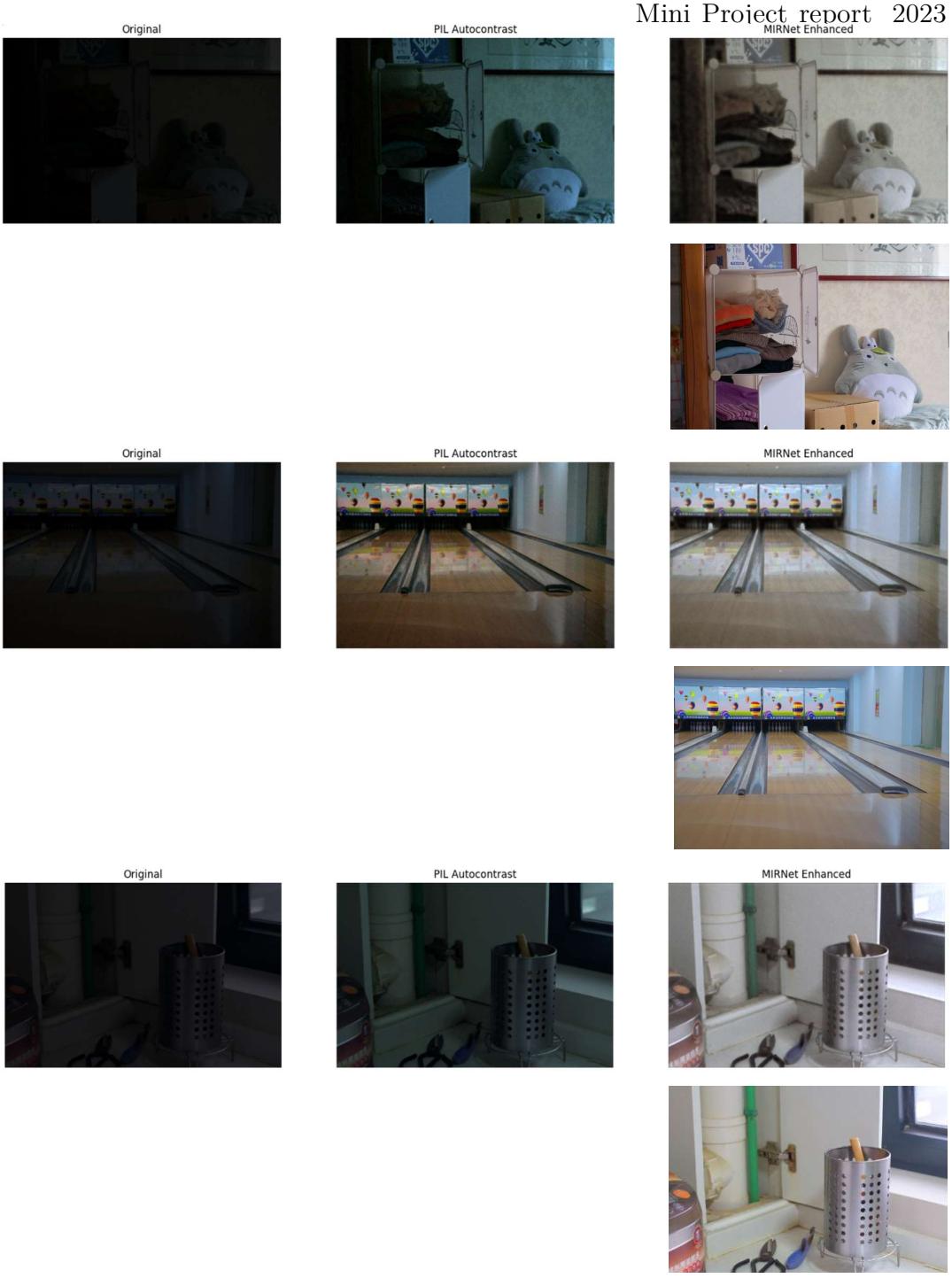


Figure 10: A Few of The Test Dataset Images From Left Original, PIL Autocontrast, MIRNet Enhanced and Below it, Ground-Truth Image

The average PSNR value out of the test dataset comes out to be: **24.14**

We have also given some of our own inputs taken with our mobile phones, one such example output is shown below

Figure 11: Image of a Hostel Room in Low Lighting With Output MIRNet Enhanced

We also compare the results of this model with other relevant models using the below cited sources and it is observed that the MIRNet produces better results than the other models using PSNR as an evaluation metric

Method	BIMEF [25]	CRM [26]	LIME [28]	MF [29]	RRM [30]	SRIE [31]	Retinex-Net [32]	MSR [33]	KinD [34]	**MIRNet (ours)**
PSNR	13.86	17.20	16.76	18.79	13.88	11.86	16.77	13.71	20.87	**24.14**

Table 1: Low Light Image Enhancement Evaluation on the LoL Dataset With Different Architectures

5 Conclusion and Future Scope

Traditional methods for image restoration and enhancement typically follow two main approaches: maintaining full-resolution features throughout the network hierarchy or adopting an encoder-decoder architecture. The former preserves precise spatial details while the latter offers improved contextualized representations. However each approach alone struggles to meet the dual requirements of real-world image restoration tasks which demand a synergistic combination of both aspects based on the input sample. In this project we implement an innovative architecture featuring a primary branch dedicated to good image processing complemented by parallel branches providing enhanced contextualized features. We make use of novel mechanisms to establish relationships among features within each branch and across multiple scale branches. This feature fusion strategy ensures dynamic adaptation of the receptive field without compromising original feature details.

This MIRNet model doesn't have to be used just for image enhancement but can also be used for image denoising and super resolution restoration purposes. Its also possible to implement this model as a real time image processing algorithm using relevant hardware such as Arduino with a camera module.

References

[1] Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: CVPR (2018)

[2] A. A. Landini and F. Rand, "Image enhancement in the spatial domain," Proceedings of the IEEE, vol. 63, no. 6, pp. 918-937, 1975.

[3] J. J. McCann, "Lightness and Retinex Theory," Journal of the Optical Society of America, vol. 61, no. 1, pp. 1-11, 1971.

[4] D. Chen and H. Zhang, "Learning a Deep Convolutional Network for Image Super-Resolution," in Proceedings of the International Conference on Computer Vision (ICCV), 2015, pp. 370-378.

[5] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure Fusion," in Proceedings of Pacific Graphics, 2007, pp. 382-390.

[6] N. Wiener, "Extrapolation, Interpolation, and Smoothing of Stationary Time Series," John Wiley & Sons, Inc., 1964.

[7] Land, E.H.: The retinex theory of color vision. Scientific american (1977)

[8] Bertalm´ıo, M., Caselles, V., Provenzi, E., Rizzi, A.: Perceptual color correction through variational techniques. TIP (2007)

[9] Palma-Amestoy, R., Provenzi, E., Bertalm´ıo, M., Caselles, V.: A perceptually inspired variational framework for color enhancement. TPAMI (2009)

[10] Jobson, D.J., Rahman, Z.u., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. TIP (1997)

[11] Ignatov, A., Timofte, R.: Ntire 2019 challenge on image enhancement: Methods and results. In: CVPRW

[12] Shen, L., Yue, Z., Feng, F., Chen, Q., Liu, S., Ma, J.: Msr-net: Low-light image enhancement using deep convolutional network. arXiv (2017)

[13] Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. BMVC (2018)

[14] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)

[15] Lore, K.G., Akintayo, A., Sarkar, S.: LLNet: a deep autoencoder approach to natural low-light image enhancement. Pattern Recognition (2017)

[16] Chen, Y.S., Wang, Y.C., Kao, M.H., Chuang, Y.Y.: Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In: CVPR (2018)

[17] Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., Van Gool, L.: Wespe: weakly supervised photo enhancer for digital cameras. In: CVPRW (2018)

[18] r, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: MICCAI (2015)

[19] Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: Deblurgan-v2: Deblurring (ordersof-magnitude) faster and better. In: ICCV (2019)

[20] Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. TPAMI (2015)

[21] Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. TIP (2017)

[22] Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: CVPR (2018)

[23] Nah, S., Kim, T.H., Lee, K.M.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: CVPR (2017)

[24] Syed Waqas Zamir, Aditya Arora "Learning Enriched Features for Real Image Restoration and Enhancement". In: *ECCV* (2020)

[25] Ying, Z., Li, G., Gao, W.: A bio-inspired multi-exposure fusion framework for low-light image enhancement. arXiv preprint arXiv:1711.00591 (2017)

[26] Ying, Z., Li, G., Ren, Y., Wang, R., Wang, W.: A new image contrast enhancement algorithm using exposure fusion framework. In: CAIP (2017)

[27] Dong, X., Wang, G., Pang, Y., Li, W., Wen, J., Meng, W., Lu, Y.: Fast efficient algorithm for enhancement of low lighting video. In: ICME (2011)

[28] Guo, X., Li, Y., Ling, H.: Lime: Low-light image enhancement via illumination map estimation. TIP (2016)

[29] Fu, X., Zeng, D., Huang, Y., Zhang, X.P., Ding, X.: A weighted variational model for simultaneous reflectance and illumination estimation. In: CVPR (2016)

[30] Liu, Y., Wang, R., Shan, S., Chen, X.: Structure inference net: Object detection using scene-level context and instance-level relationships. In: CVPR (2018)

[31] Jobson, D.J., Rahman, Z.u., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. TIP (1997)

[32] ang, S., Zheng, J., Hu, H.M., Li, B.: Naturalness preserved enhancement

algorithm for non-uniform illumination images. TIP (2013)

[33] Wang, W., Wei, C., Yang, W., Liu, J.: Gladnet: Low-light enhancement network with global awareness. In: FG (2018)

[34] Zhang, Y., Zhang, J., Guo, X.: Kindling the darkness: A practical low-light image enhancer. In: MM (2019)