CS512 Computer Vision – Project Report

LYT-Net: Lightweight YUV Transformer-based Network for Low-Light Image Enhancement

Team Members

- Tamilarasee Sethuraj | A20553416
- Kavin Raj Karuppusamy Ramasamy | A20564249

Research Paper

Title	:	LYT-Net: Lightweight YUV Transformer-based Network for Low-Light Image Enhancement
Authors	:	Alexandru Brateanu, Raul Balmez, Adrian Avram, Ciprian Orhei, and Cosmin Ancuti
Year	:	2024
Link	:	https://arxiv.org/abs/2401.15204

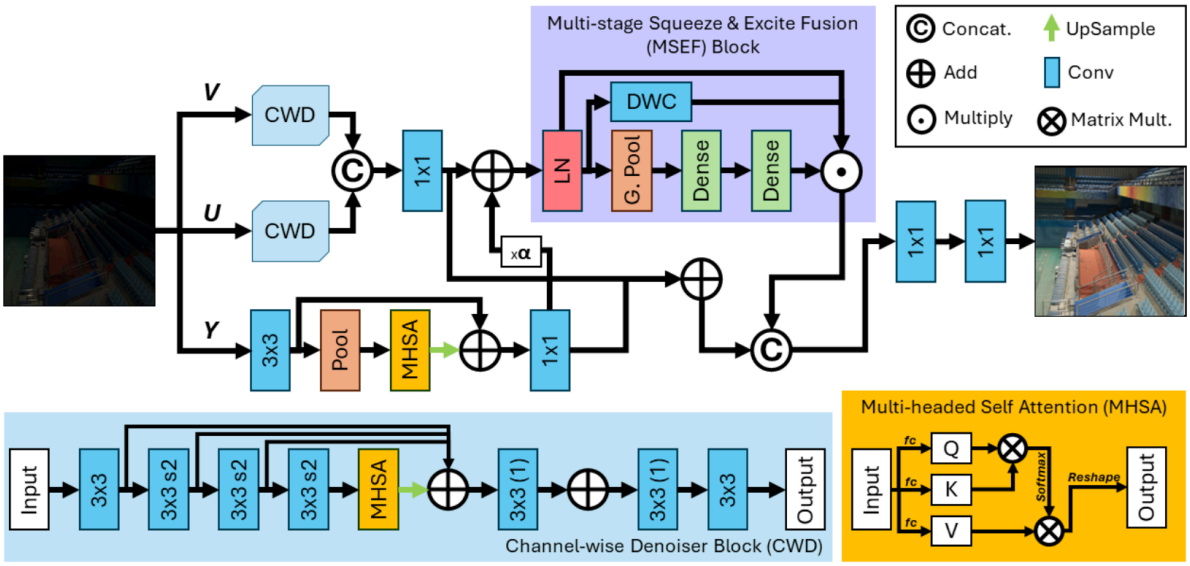
Problem Statement

Low-light image enhancement (LLIE) is a crucial task in computer vision, aimed at improving the quality of images captured under suboptimal lighting conditions. Images taken in low light often suffer from issues like low contrast, high noise, color distortion, and loss of detail, all of which degrade their quality and hinder the performance of downstream vision tasks such as object detection, recognition, and segmentation. Enhancing these images is essential in applications like surveillance, autonomous driving, and medical imaging, where visibility and clarity are critical for accurate analysis. However, LLIE presents several challenges, including amplifying noise, maintaining color fidelity, and preserving fine details without introducing artifacts. Additionally, many existing deep learning-based approaches, while effective, are computationally expensive and unsuitable for real-time applications.

Traditional methods like histogram equalization and Retinex-based models, though widely used, often fail to meet the demands of complex lighting scenarios and tend to introduce noise or distort color balance. Recent deep learning approaches, including CNN-based models like Retinex-Net and EnlightenGAN, have shown promising results but are often computationally intensive. More efficient models, such as transformer-based architectures like Uformer and Restormer, leverage self-attention mechanisms to capture global dependencies in images but still require significant computational resources. Given the increasing demand for real-time LLIE solutions, especially in mobile and embedded systems, there is a need for lightweight models that balance performance with efficiency. This project implements LYT-Net, a lightweight

transformer-based model designed to enhance low-light images by processing them in the YUV color space, achieving state-of-the-art results with minimal computational overhead.

Proposed Solution



The LYT-Net architecture is a lightweight, transformer-based model designed specifically for Low-Light Image Enhancement (LLIE). It efficiently processes images in the YUV color space, where the luminance (Y) and chrominance (U and V) channels are treated separately. This approach allows the model to focus on adjusting illumination and removing noise more effectively while preserving color fidelity. LYT-Net comprises several critical components, including:

1. Channel-Wise Denoiser (CWD)
2. Multi-Headed Self-Attention (MHSA)
3. Multi-Stage Squeeze & Excite Fusion (MSEF)

The LYT-Net architecture adopts a dual-path design that separates the Y (luminance) channel from the U and V (chrominance) channels. This separation allows the network to more effectively handle illumination corrections and reduce noise in low-light images. The Y channel, responsible for brightness and illumination, is processed with attention mechanisms to adjust

lighting, while the U and V channels, which carry color information, are denoised using specialized blocks.

The architecture is built with efficiency in mind, making use of lightweight transformer blocks to ensure that the model can run on resource-constrained devices without sacrificing performance.

YUV Model:

The YUV format is a color encoding system that separates an image into luminance (Y) and chrominance (U and V) components, designed for efficient image processing and compression. The Y component represents the luminance or brightness, capturing the overall light intensity of the image, while U and V represent chrominance, encoding color information by describing the differences between the blue and red channels, respectively. In contrast, the RGB model represents images based on the intensities of red, green, and blue at each pixel, where all three channels contribute equally to both color and brightness. YUV is advantageous because it aligns more closely with human visual perception by prioritizing brightness (to which the human eye is more sensitive) over color. This separation allows for more efficient compression, with minimal loss of quality, as chrominance data can often be downsampled without significant perceptual degradation. In LLIE tasks like the LYT-Net implementation, YUV is particularly useful because it allows independent processing of brightness (Y) and color (U and V), enabling targeted enhancements to luminance while minimizing color distortions and improving computational efficiency.

Channel-Wise Denoiser (CWD)

The Channel-Wise Denoiser (CWD) is designed to reduce noise in the chrominance channels (U and V) while ensuring that fine image details are preserved. It employs a U-shaped architecture with multiple convolutional layers and a Multi-Headed Self-Attention (MHSA) mechanism at the bottleneck. The U-shaped structure is ideal for low-level vision tasks, as it captures multi-scale features and uses skip connections to prevent the loss of fine details during upsampling.

The CWD consists of four convolutional layers:

- The first layer processes features with a stride of 1 to extract initial spatial features.

- The next three layers use strides of 2 to downsample the feature map, progressively capturing features at multiple scales.

- At the bottleneck, the MHSA mechanism captures long-range dependencies within the image, enhancing its ability to focus on important regions of the chrominance channels.

- After the MHSA block, the feature map is upsampled using interpolation-based methods (instead of transposed convolutions), reducing the number of parameters while maintaining performance.

By applying the CWD block to the U and V channels, LYT-Net effectively reduces noise while retaining important color information. This separation allows for more focused noise reduction, avoiding the risk of over-processing the luminance channel, which is critical for maintaining natural lighting in the enhanced image.

Multi-Headed Self-Attention (MHSA)

The Multi-Headed Self-Attention (MHSA) block is applied to the Y (luminance) channel, which is responsible for handling contrast and illumination. The MHSA block is inspired by the Vision Transformer (ViT) architecture and is designed to capture long-range dependencies between different regions of the image. This is crucial for adjusting lighting in low-light images, where global context plays an important role in enhancing poorly lit areas.

In the MHSA block:

- The input feature map is projected into query (Q), key (K), and value (V) matrices using linear layers.

- The self-attention mechanism is applied across multiple heads, enabling the model to attend to different parts of the image simultaneously.

- The attention outputs from all heads are concatenated and passed through a final linear layer to restore the original feature size.

Mathematically, the self-attention mechanism is defined as:

$$Attention(Q,\ K,\ V)\ =\ Softmax(\frac{QK^T}{\sqrt{d_k}})\ V$$

where d_k is the dimensionality of the key vectors. The softmax operation ensures that the model focuses on the most relevant parts of the image for enhancing illumination.

Once the self-attention mechanism has been applied, the output is reshaped back to its original dimensions and passed to the next stage. The MHSA block allows LYT-Net to adjust illumination while preserving fine details, ensuring that the enhanced image looks natural without introducing artifacts.

Multi-Stage Squeeze & Excite Fusion (MSEF)

The Multi-Stage Squeeze & Excite Fusion (MSEF) block is applied after the luminance (Y) and chrominance (U, V) channels have been processed separately. The MSEF block enhances both spatial and channel-wise features by employing a combination of global average pooling and excitation mechanisms.

The squeeze operation compresses the feature map into a reduced set of descriptors that capture global context. This operation is followed by an excitation step, which re-expands the descriptors to their original dimensions while selectively emphasizing the most important features. This process helps the network focus on the most relevant spatial and channel-wise details, improving the overall quality of the enhanced image.

The MSEF block works as follows:

1. The input feature map undergoes layer normalization to stabilize the learning process.

2. Global average pooling is applied to capture the global spatial context.

3. The pooled features are passed through fully connected layers with ReLU and Tanh activations, which produce a descriptor that emphasizes important features.

4. The feature map is then re-expanded to its original dimensions and recombined with the input feature map using a residual connection.

This approach allows the model to enhance both local and global features, ensuring that the final image is not only well-lit but also retains important structural and color details.

Hybrid Loss Function

To train the LYT-Net model, a hybrid loss function is employed, combining several loss components to ensure that the enhanced images are both perceptually pleasing and quantitatively accurate. The hybrid loss function is defined as:

$$L = L_S + \alpha_1 L_{prec} + \alpha_2 L_{hist} + \alpha_3 L_{psnr} + \alpha_4 L_{color} + \alpha_5 L_{MS-SSIM}$$

Where:

- L_S : Smooth L1 loss, which helps minimize pixel-wise errors between the target and predicted images.

- L_{prec} : Perceptual loss, which ensures that high-level features extracted by a pre-trained VGG19 network are similar between the target and predicted images.

- L_{hist} : Histogram loss, which aligns the pixel intensity distributions of the enhanced and target images, ensuring proper contrast.

- L_{psnr} : PSNR loss, which penalizes reconstruction errors based on the Peak Signal-to-Noise Ratio (PSNR).

- L_{color} : Color loss, which ensures that the color balance of the enhanced image matches that of the target.

- $L_{MS-SSIM}$: Multi-Scale SSIM loss, which evaluates structural similarity across multiple scales, ensuring that both local and global structures are preserved.

This hybrid loss function ensures that the model optimizes various aspects of image quality, leading to enhanced images that are visually pleasing and maintain high structural and color fidelity.

Data:

The datasets used in this project for training and evaluating the LYT-Net model are the LOL-v1 and LOL-v2 datasets, which are widely used for Low-Light Image Enhancement (LLIE) tasks. The LOL-v1 dataset consists of paired low-light and normal-light images, specifically designed for enhancing images captured in dim environments. It contains a training set of 485 low-light images paired with their corresponding 485 normal-light images, as well as a test set with 15 low-light and 15 normal-light images. This dataset is relatively small but provides a controlled

environment for evaluating the effectiveness of image enhancement techniques. The images are organized into separate folders for input (low-light) and target (normal-light) images, and they are preprocessed using techniques like normalization and augmentation to improve model performance.

The LOL-v2 dataset is an expanded version of LOL-v1 and includes two distinct subsets: LOL-v2-real and LOL-v2-synthetic. The LOL-v2-real subset consists of 689 low-light images paired with 689 normal-light images for training, while the test set includes 100 paired images. These images are captured in real-world low-light scenarios, making the dataset more challenging and diverse. On the other hand, the LOL-v2-synthetic subset includes 900 training images and 100 test images, which are synthetically generated from normal-light images to simulate low-light conditions. This subset is useful for testing the model's generalization to artificially created low-light environments. Both datasets provide a comprehensive evaluation framework for LYT-Net, enabling it to be tested across a variety of real and synthetic low-light scenarios.

Link: https://daooshee.github.io/BMVC2018website

Team Member Responsibilities (in %)

Task	Tamilarasee	Kavin
Data Preprocessing and Augmentation	50	50
Model Implementation	50	50
Evaluation and Analysis	50	50
Experimentation and Improvement	50	50
Report Writing	50	50
Presentation Preparation	50	50

References:

- C. Wei, W. Wang, W. Yang, and J. Liu, “Deep retinex decomposition for low-light enhancement”, in Proceedings of the British Machine Vision Conference (BMVC), 2018
- S. Park, S.Yu, B.Moon, S.Ko, and J. Paik, “Low-light image enhancement using variational optimization-based retinex model”, IEEE Transactions on Consumer Electronics, vol. 63(2), 2017.
- Shansi Zhang, Nan Meng and Edmund Y. Lam, “LRT: An Efficient Low-Light Restoration Transformer for Dark Light Field Images”, IEEE Transactions on Image Processing, vol. 32, 2023.