Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.059669

ARTICLE

A Transformer Network Combing CBAM for Low-Light Image Enhancement

Zhefeng Sun* **and Chen Wang**

The Center for Information of National Medical Products Administration, Beijing, 100076, China

*Corresponding Author: Zhefeng Sun. Email: sunzf@nmpaic.org.cn

Received: 14 October 2024; Accepted: 03 January 2025; Published: 06 March 2025

ABSTRACT: Recently, a multitude of techniques that fuse deep learning with Retinex theory have been utilized in the field of low-light image enhancement, yielding remarkable outcomes. Due to the intricate nature of imaging scenarios, including fluctuating noise levels and unpredictable environmental elements, these techniques do not fully resolve these challenges. We introduce an innovative strategy that builds upon Retinex theory and integrates a novel deep network architecture, merging the Convolutional Block Attention Module (CBAM) with the Transformer. Our model is capable of detecting more prominent features across both channel and spatial domains. We have conducted extensive experiments across several datasets, namely LOLv1, LOLv2-real, and LOLv2-sync. The results show that our approach surpasses other methods when evaluated against critical metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Moreover, we have visually assessed images enhanced by various techniques and utilized visual metrics like LPIPS for comparison, and the experimental data clearly demonstrate that our approach excels visually over other methods as well.

KEYWORDS: Low-light image enhancement; CBAM; transformer

1 Introduction

Low-light image enhancement involves the technique of boosting the clarity and visual quality of photographs captured under insufficient lighting conditions. It encompasses various tasks, including noise reduction, color correction, and detail enhancement. Many algorithms have been proposed in these topics by researchers.

These methods can be categorized into three types, namely traditional methods, CNN-based methods, and transformer-based methods, but each type has its drawbacks. Traditional methods like histogram equalization, gamma correction, contrast stretching, and Contrast Limited Adaptive Histogram Equalization (CLAHE), aim to adjust in the statistical variables of the image to obtain a more reasonably distributed picture. However, these methods lack an understanding of the real physical world image generation process, which significantly limits their effectiveness. The introduction of the Retinex theory has modeled the process by which real-world observers obtain images well. Therefore, a lot of subsequent research has been proposed based on this theory, including algorithms such as SSR, MSR. The Retinex theory suggests that an image's appearance is shaped by the combination of the light that illuminates the scene and the way objects within the scene reflect that light. A multitude of methods under this theoretical framework tend to focus on illumination estimation and filtering transformations. However, the real world often contains various types of noise and random factors, making the image generation process more complex than the theory assumes

and limiting the performance of these methods. Besides, these types of methods are often analytical methods and do not use labeled data like modern machine learning methods.

As deep learning technology has advanced and been applied across numerous domains in recent years, it has also been harnessed for the purpose of enhancing images captured in low-light conditions. A plethora of deep learning-based methods have been developed for this specific application. The researchers design different network structures for these tasks and then learn the corresponding knowledge from labeled data through offline supervised learning. Then, the well-learned model is used to process new low-quality images. The typical network structure is Convolutional Neural Networks (CNNs). This method usually uses several CNNs to learn aspects such as color, denoising, and illumination independently and then joint learning in sequence to obtain the final model. However, CNN often focuses on the local spatial properties of images and cannot handle non-local similarities and long-range dependency issues very well.

Therefore, some scholars have now proposed network structures based on transformers to overcome the issues related to CNNs. The current state-of-the-art (SOTA) method, RetinexFormer, is one of such approaches. It proposes a one-stage joint training method called ORF (One-stage Retinex-based Framework) and introduces IGT (Illumination-Guided Transformer) structure to deal with long-range dependencies and non-local similarities, while also integrating the IGT into the ORF. It shows that this method currently achieves the best results in several dataset.

However, due to the extremely complex conditions of real-world imaging, coupled with inherent challenges in the Transformer architecture, such as issues related to positional encoding, we argue that incorporating additional attention modules of other dimensions prior to the Transformer processing can enforce the model focus on important features. Therefore, we propose a method that integrates CBAM with the Transformer. Prior to Transformer processing, the CBAM applies dual-dimensional attention, subsequently delivering the refined outputs to the Transformer modules. This enhanced attention mechanism enables the Transformer to receive more informative inputs, leading to improved performance. Our contributions can be summarized as follows:

1. We introduce a novel method that integrates CBAM with the Transformer for low-light image enhancement. This approach effectively prioritizes significant features prior to processing by the Transformer, leading to improved performance.
2. Through extensive experiments on widely used and publicly available benchmark datasets, we have shown that our method achieves significant improvements over existing methods in key performance metrics.

This paper is organized as follows:

The first part is a brief introduction, which provides basic cognition, our motivation, and the highlights of our work.

The second part discusses related work, including traditional image enhancement methods, CNN-based image enhancement methods, transformer-based image enhancement methods, and attention mechanism.

The third part mainly describes our approach. We start with an introduction to the background knowledge, then present our network architecture, and finally provide a detailed explanation of the CBAM module.

The fourth part focuses on our experimental results. In this part, we introduce our dataset and evaluation metrics, followed by the details of our experiments, including specific parameter settings. Lastly, we compare our experimental results with other methods.

The fifth part summarizes the paper.

2 Related Work

Low-light image enhancement aims to enhance the clarity and visual quality of images captured in environments with insufficient lighting. Subtasks such as color correction, noise reduction and detail enhancement have been extensively explored, leading to the development of various algorithms. These approaches can generally be divided into three main categories, each with inherent limitations.

2.1 Traditional Methods

Traditional methods, such as histogram equalization, gamma correction, contrast stretching, CLAHE, adjust the statistical variables of the image to obtain a more reasonably distributed picture. Yet, these methods lack an understanding of the real physical world image generation process, which significantly limits the effectiveness of these methods.

Retinex theory models the real physical world image generation process and decomposes an image into two components, the illumination component and the reflectance component. A multitude of methods under this theoretical framework tend to focus on illumination estimation. For example, Jobson proposed the MSR and MSRCR algorithms, aiming to improve image enhancement effects, especially in dealing with illumination changes and color constancy. The MSRCR algorithm further addresses the potential color distortion issues of the MSR algorithm by incorporating a color restoration factor to regulate the color equilibrium within the image. However, the real world often contains various types of noise and random factors, making the image generation process more complex than the theory assumes.

2.2 CNN-Based Methods

As the deep learning revolution led by Hinton, CNN models have been applied to the domain of image enhancement for low-light conditions.

Chen et al. [1] improved low-light image quality with a fully convolutional network. Lv et al. [2] introduced MBLLEN for multi-level feature capture. Ignatov et al. [3] used a residual CNN for image enhancement. Lore et al. [4] designed a deep autoencoder for signal feature extraction. Gharbi et al. [5] presented a bilateral grid-inspired network. Cai et al. [6] trained a CNN for SICE. Guo et al. [7] created Zero-DCE, a model for dynamic range adjustment. Moran et al. [8] used deep learning for filter-based image enhancement. Xu et al. [9] proposed to reinforce image edges through structural modeling using U-Net, while employing a structure-guided enhancement module to improve the performance. Wu et al. [10] proposed a novel framework by utilizing semantic information to improve the performance.

CNN models also been applied combines with Reitnex theory. Wei et al. [11] introduced a novel approach called Retinex-Net, which incorporates Decom-Net and Enhance-Net. Wang et al. [12] introduced GLADNet for calculating and adjusting global illumination. Zhang et al. [13] developed KinD network for efficient light adjustments. Wang et al. [14] crafted a neural network to boost the quality of underexposed images by leveraging intermediate lighting cues.

Some researchers have also proposed many unsupervised learning and semi-supervised methods based on the structure of CNN. Jiang et al. [15] developed EnlightenGAN, an unsupervised GAN for image enhancement without the need for paired training data. Yang et al. [16] used a semi-supervised DRBN for the task, refining representations through adversarial learning with unpaired data. Liu et al. [17] introduced RUAS, while Wu et al. [18] presented URetinex-Net to deal with this task. Fu et al. [19] created PairLIE, an unsupervised method for training the model from low-light image pairs. Wang et al. [20] designed a reversible network for aligning the distribution of exposed images with a Gaussian model. Ye et al. [21] enhanced low-light images with EMNet, incorporating external memory into their architecture. Fei et al. [22]

applied the Generative Diffusion method as a computational prior, which improve the performance of this task.

Some researchers have proposed the fast and robust algorithms for this task. Ma et al. [23] designed SCI, a framework for quickly brightening low-light images with adaptability and robustness. Zeng et al. [24] enhanced photos rapidly and reliably using custom 3D LUTs.

However, CNN based methods show the limitations in non-local similarities and long-range dependency.

2.3 Transformer-Based Methods

Transformer is proposed by Ashish et al. to solve machine translation task, and widely be used in various areas. Some researchers also used this method for low-light image enhancement.

Chen et al. [25] proposed the IPT model, a pre-trained model using multi-head training with contrastive learning for diverse image processing tasks. Xu et al. [26] presented a method using Signal-to-Noise Ratio-aware transformers for dynamic pixel enhancement. Cui et al. [27] proposed IAT, a network for reconstructing RGB images from low-light or extreme exposure conditions. Wang et al. [28] introduced Uformer for image restoration with a hierarchical structure. Fu et al. [29] created LEGAN, an unsupervised GAN-based network trained on unpaired images. Zamir et al. [30] enhanced a transformer model with key design improvements for long-range pixel interaction capture.

Wang et al. [31] introduced LLFormer for low-light image enhancement. Dang et al. [32] presented the WaveNet, which excels in various parameters and enhances feature representation through a wave-like feature representation. Wang et al. [33] proposed FourLLIE, a two-stage Fourier-transform-based network that uses transformer technology. Wang et al. [34] suggested a method that combines the effectiveness of gamma correction with deep learning networks' robust modeling capabilities, allowing the gamma correction factor to be adaptively learned through a progressive perception of illumination deviations. Liu et al. [35] proposed the brightness-aware attention mechanisms for this task.

Yi et al. [36] proposed physically interpretable generative diffusion model along with a transformer network architecture. Dang et al. [37] introduced PPformer, which has cross-attention mechanisms in pixel-wise and patch-wise. Cai et al. [38] proposed an efficient model that estimates illumination to brighten low-light images, followed by corruption restoration for enhancement.

2.4 Attention Mechanism

The attention mechanism mimics a cognitive process in human visual perception. It allows individuals to swiftly survey an entire scene to identify areas of interest. Subsequently, these areas are allocated more cognitive resources to extract details, while irrelevant information is filtered out. Xu et al. [39] proposed and apply this mechanism in CV areas, and a lot of methods follow this work and get good result in CV tasks. Vaswani et al. [40] proposed the self-attention in NLP tasks and achieving groundbreaking progress. Zhou et al. [41] proposed the DIN methods and apply the attention mechanism in Advertising and get significant revenue improvement for big company. More closely to our work, Woo et al. [42] proposed CBAM, which uses dual dimensional attention method and gets better result.

3 Method

3.1 Preliminary

Retinex theory is a theory that explains human color constancy and the mechanism for computing lightness values from an image. It suggests that the human visual system perceives color relatively independently

of the illumination conditions. The theory is based on the idea that the color sensation comes from the ratio of light reflected from different surfaces, rather than the absolute amount of light.

In the context of image processing, we can separate an image into illumination components and reflectance components based on Retinex theory. The foundational model posits that any specific image is the result of pixel-wise multiplication between two images, the illumination image (L) and the reflectance image (R). Once obtaining the illumination image, the reflectance image, which contains the intrinsic details of the image, can be derived. Retinex algorithms are used in various tasks such as low-light enhancement, dynamic range compression, and single image dehazing. They have the advantage of being relatively easy to implement.

However, these methods also have shortcomings, such as color distortion, local halo effects, and loss of local details. Color distortion refers to the alteration or degradation of the true colors of an image or visual representation. One of the primary issues with Retinex algorithms is the introduction of color artifacts, particularly around edges and highlights. This is due to the method's inability to accurately separate the reflectance and illumination components of an image, especially when there are rapid changes in color or brightness.

Local halo effects refer to an artifact that can occur in image processing, particularly in techniques such as the Retinex algorithm. These effects manifest as unwanted bright or dark rings or fringes that appear around objects or edges in an image. They are called "halos" because they resemble the halos seen around bright objects in low-contrast images. The local halo effect typically happens when an image enhancement algorithm attempts to adjust the brightness or contrast of an image. The algorithm may not accurately preserve the local contrast around edges, leading to a phenomenon where the edges appear to have a glowing or halo-like appearance. This is often caused by the blurring or smoothing operation that is applied to the image during processing, which can bleed lightness or darkness from the edges into the surrounding areas, creating a halo. For example, when using a Gaussian blur to estimate the image's illumination component, the sharp transitions at edges may be smoothed too much, causing the halo effect. This effect can be visually distracting and is generally considered undesirable because it does not represent the true detail of the original image.

Another drawback of the Retinex theory is that it can lead to local details loss. There may be several reasons. The first reason is that the Gaussian filter or other smoothing operations used in Retinex algorithms to estimate the illumination component can cause over-smoothing, which may result in the blurring of edges and a loss of fine details. The second reason is that noise is often a problem in low-light images. While attempting to enhance illumination, Retinex algorithms may inadvertently amplify noise, which can obscure fine details in the image. The third reason is the theory's assumptions about the separation of illumination and reflectance might not hold in all scenarios, particularly in scenes with complex lighting or where the light sources are not well understood. It may result in suboptimal decomposition, which consequently lead to a loss of detail.

According to the Retinex theory and corruptions in the real world, we can decompose a low-light image I into two parts, a reflectance image R and an illumination map L,

$$I = R \cdot L, \tag{1}$$

In which I is a low-light image, assumed to have a height of H and a width of W, with three RGB channels, thus I is a tensor of size 3*H*W. R is the reflected image, is also a tensor of size 3*H*W. L is the illumination prior, which is a matrix of size H*W. The multiplication in the formula is element-wise multiplication.

However, the real world does experience corruption, which may be due to high ISO and long exposure settings in dark scenes, or the brightening process that may amplify noise and artifacts. Therefore, we need to revise this formula. We introduce a perturbation term for both R and L to represent this corruption, as

$$I = (R + R') \cdot (L + L') = R \cdot L + R \cdot L' + R' \cdot (L + L'). \tag{2}$$

In which R' is the perturbation of the reflectance image R, and R' is also a tensor of size 3*H*W. L' is the illumination perturbation, which is a matrix of size H*W.

Following the conclusions discussed in [14,43,44], it is assumed that R is a well-exposed image. So, it can multiply a light-up map L'' to light up I and which L'' * L = 1 as

$$I \cdot L'' = R + R \cdot (L' \cdot L'') + (R' \cdot (L + L')) \cdot L''. \tag{3}$$

In which $R \cdot (L' \cdot L'')$ can indicates the exposure imbalance and color distortion, $(R' \cdot (L + L')) \cdot L''$ represents the noise. We can simplify it as

$$I_{lu} = I \cdot L'' = R + C, \tag{4}$$

where C is the overall corruption term, and I_{lu} is a tensor of size 3*H*W which represent the lit-up image. So, we can formulate the Framework as

$$(I_{lu}, F_{lu},) = f(I, L_p) \qquad I_{en} = R\,(I_{lu}, F_{lu})\,, \tag{5}$$

where L_p is illumination prior matrix of size H*W and can get calculates the mean values along the channel dimension of the origin image I, as

$$L_p = mean_{channel}(I), \tag{6}$$

where f represents the function of an illumination estimator which takes I and L_p as inputs. It outputs the light-up feature F_{lu} which is a tensor of size 3*H*W and lit-up image I_{lu}. In our work, we will simulate this function using a simple neural network as Fig. 1. The detailed network design will be introduced in the next section.

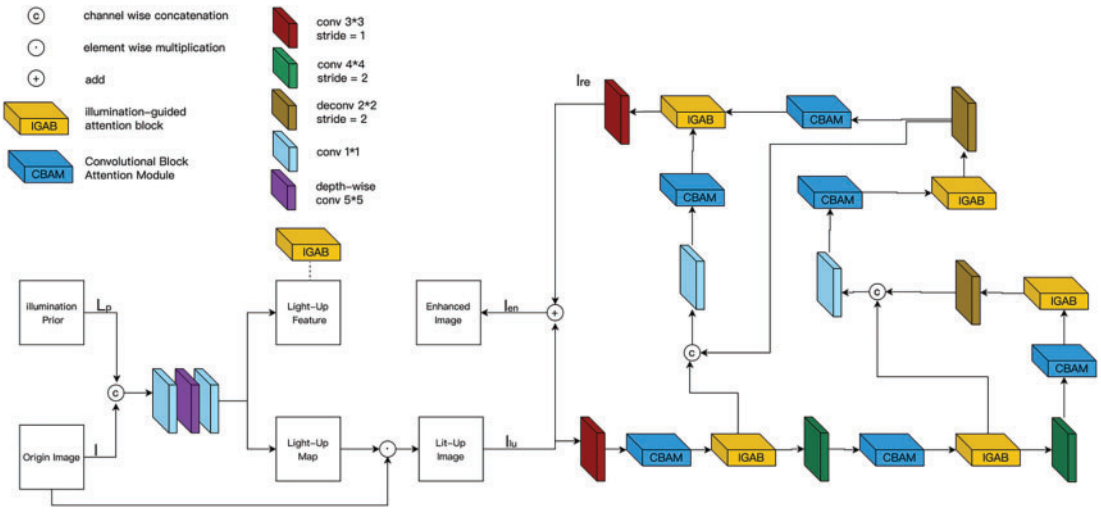

Figure 1: This shows the entire design framework. The entire data pipeline consists of two parts, illumination estimation and corruption restorer. We represent the latter with a collaborative network formed by convolutional networks, CBAM, and Illumination-Guided Attention Transformer (IGAT)

R is a function denotes the corruption restorer and it takes the F_{lu} and I_{lu} as inputs. It outputs the final enhanced image I_{en}. In our work, we will simulate this function using a series of complex network structures such as CNNs, IGAB, and CBAM. The detailed network design will be introduced in the next section.

L_1 loss function is our loss function to guide the learning of the entire model parameters. Assuming I and I′ are the ground truth image and the predicted image, and I_{p} and I'_{p} are the pixels of the ground truth and predicted images, and n is the total number of pixels, respectively, then the L_1 loss can be expressed as

$$L_1 = \sum_p^n |I_p - I'_p|. \tag{7}$$

We employ the traditional backpropagation algorithm to propagate the gradients backward, to update the corresponding parameters. After several rounds of iteration, we obtain our final model. This is the process of loss-guided model training.

3.2 Network Architecture

The input of the whole network is original image and illumination prior. Illumination prior refers to the assumptions or knowledge about the lighting conditions of a scene that can be used to improve the performance. This can include the understanding that certain areas of an image should be brighter or that the lighting is coming from a particular direction. By using these priors, algorithms can make educated guesses about how to adjust the image to compensate for poor lighting conditions and enhance the overall image quality. In our algorithm, we will use illumination prior as one of the input data sources.

As Fig. 1 shows, the entire neural network can be divided into two sub-parts, illumination estimation and corruption restorer.

Firstly, we concatenate the image I and illumination prior Lp by channel wise, this operation aims to enhance the model's expressive power. Then we incorporate the obtained results into a preceding neural network. In this network, we use a convolutional neural network with a 1 × 1 kernel to fuse the results for the first time. Then we use a depth-wise separable conv 5 × 5 to model the interactions of regions with different lighting conditions and generate the Flu. This operation aims to use the well-exposed regions as semantic contextual information for the under-exposed regions. The depth-wise separable convolution breaks down the standard convolution into two simpler operations. The first part is depth-wise convolution. Each input channel is convolved with a separate kernel, producing outputs for each input channel individually. If the input has N channels, then N separate filters are applied, one for each input channel. The second part is point-wise convolution. After the depth-wise step, a 1 × 1 convolution is used to combine the outputs from the depth-wise step. This step mixes the information from the depth-wise pass to create the final output channels. The point-wise convolution can significantly reduce the number of parameters because it uses small filters that slide across the input feature map. Then we use a conv 1 × 1 to aggregate the Flu to produce the light-up map L. Then we combine the light-up map with origin image by element wise multiplication. This nonlinear operation primarily serves to partially enhance features. Finally, we get the lit-up image, and the obtained results are used as inputs for the second part.

For the second part, we employ a deep network with a complex structure to simulate the function R. The basic units of this network structure are the convolutional layer, CBAM, and IGAB.

The convolutional layer primarily extracts locally similar features. This layer applies a set of learnable filters to the input data. These filters slide across the input image, computing the dot product between the image pixels and the filter weights, which results in a new set of features or feature maps. Meanwhile, we also use different strides in different convolutional layers. Stride refers to the step size the filter (or kernel) takes as it moves across the input image. The stride determines how the convolution operation progresses and has

a significant impact on the output dimensions of the feature map. In this work, we often set the stride value equals 1 or 2.

CBAM focuses on enhancing attention in dual dimensions. CBAM sequentially generates channel attention map and spatial attention map from the input data of the module. More details will be introduced in the next part.

IGAB is a transformer-inclusive network structure that deeply processes feature structures. IGAB first performs layer normalization on the input data I and then puts the output into the IGMSA module. The IGMSA module refers to Illumination-Guided Multi-head Self-Attention, which adds an illumination guide item to the original Multi-head Self-Attention to guide the entire process. The obtained results are then added to the original image. The operation is similar to the mechanism of a Residual Network (ResNet). Skip connections allow gradients to flow directly through layers, alleviating the vanishing gradient problem. We do the layer normalization again for the result II. Layer normalization operates by normalizing the inputs to a layer such that for each input vector, the mean and variance are adjusted to 0 and 1, respectively. This is done independently for each input vector, making it particularly useful for sequences of varying lengths. Then we pass the data through a Feed-Forward Network (FFN) layer. Finally, we add the result to II and get the result. Fig. 2 depicts the entire workflow of IGAB.

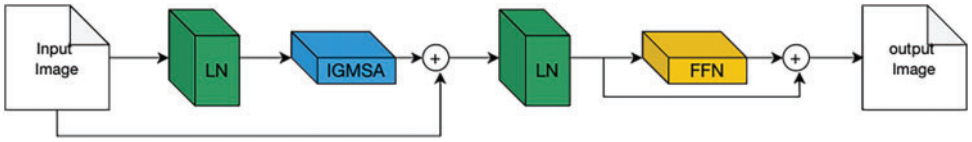

Figure 2: The whole process of IGAB

We allow the input features to sequentially pass through these three modules multiple times. First, we process the data with a convolutional layer which the kernel is 3 × 3 and the stride is 1. After that, we will use CBAM to process the data. Then we use IGAB dealing with the data. We repeat the process several times. After the operations as shown in Fig. 1, we obtain the feature map Ire. It is then added to Ilu to get the final enhanced feature map.

3.3 CBAM

CBAM successively generates a one-dimensional channel attention map followed by a two-dimensional spatial attention map, both derived from an input feature map. Channel attention module determines which channels (or feature maps) are more important. It utilizes pooling techniques, including global average pooling and global max pooling, to generate two separate channel descriptors, which are combined and passed through a couple of convolutions and a sigmoid layer to produce channel attention maps. After the channel attention has been applied, the spatial attention module focuses on which spatial locations are more informative. It uses the input that has already been refined by the channel attention and applies a convolution to generate spatial attention maps. Figs. 3 and 4 show the whole process of CBAM.

CBAM's Channel Attention Module starts by collecting spatial details from the input feature map using average and max pooling. These operations yield two separate spatial context descriptors that are subsequently refined by a shared network to form a channel attention map. This network is generally made up of a multi-layer perceptron (MLP) featuring a hidden layer. The MLP outputs for each descriptor are combined through an element-wise summation to generate the channel attention map. Finally, a sigmoid function is applied to the channel attention map to scale the values.

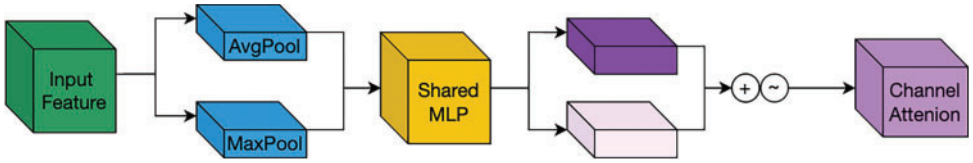

Figure 3: Channel attention

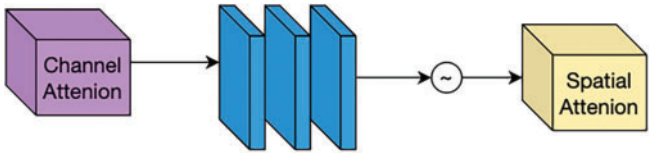

Figure 4: Spatial attention

CBAM's Spatial Attention Module is designed to concentrate on the critical spatial regions within the input feature map. The process involves subjecting the input feature map to both global average pooling and global max pooling, which results in two unique feature maps, with each map highlighting a different aspect of spatial information. Then these two feature maps are then concatenated to form a single feature map that combines both the average and max pooled features. The concatenated feature map is passed through a convolutional layer which helps to refine the spatial attention. The output of the convolutional layer is then passed through a sigmoid activation function to generate the final spatial attention map. This map highlights areas of the feature map which are considered important.

4 Experiment

4.1 Datasets and Metrics

Our approach is assessed using datasets from the Low-Light (LOL) benchmark, which consists of two versions: v1 and v2. Version 2 of LOL is further segmented into synthetic and real parts. The allocation for training and testing within these datasets is as follows: for LOL-v1, it's 485 to 15, for the real subset of LOL-v2, it's 689 to 100 and for the synthetic subset of LOL-v2, it's 900 to 100.

We use PSNR and SSIM as the primary evaluation metrics to assess our algorithm. PSNR is a widely used metric in the field of image and video processing to quantify the quality of a reconstructed image or video compared to its original version. It measures the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. In simpler terms, PSNR is a way to numerically express the similarity between the compressed image and the original image. A higher PSNR value indicates a higher quality image, with less noise and artifacts.

SSIM is a metric used in the field of image processing to measure the quality of an image, particularly in terms of its structural degradation. It compares the structural information between a reference image (original) and a distorted image (processed or compressed) to determine the level of similarity. SSIM is designed to reflect the perceived change in structural information or the loss of correlation between the structures in the original and the distorted image. It considers three components: luminance, contrast, and structure. The SSIM index is calculated based on these components and ranges from −1 to 1, where 1 indicates perfect similarity.

4.2 Implementation Details

The model is trained by the PyTorch framework and use an adaptive momentum estimation (Adam) optimizer. We train our model for 1.5×10^5 iterations. We set the learning rate as 2×10^{-4} initially. After several rounds of training, we reduced the learning rate to 1×10^{-6} using a cosine annealing scheme. We configured the batch size as 8 and randomly crop the images into patches of size 384×384 as training inputs and use random rotation and flipping for data augmentation. We conduct experiments on one RTX 3090 GPU.

4.3 Comparison with State-of-the-Art Methods

First, we will conduct a quantitative experimental analysis. Since paper provides a detailed comparison of experimental metrics for dozens of current popular methods, we will excerpt its experimental data and adopt its conclusion that Retinexformer is the SOTA method across several datasets. Then, we will focus on comparing the results of Retinexformer with our results. We combined the results from the paper with our experimental results to obtain Table 1.

Table 1: Comparisons on the LOL dataset excerpt from the Retinexformer Paper

Methods	LOL v1		LOL v2-real		LOL v2-syn	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
3DLUT [24]	14.35	0.445	17.59	0.721	18.04	0.8
DeepUPE [14]	14.38	0.446	13.27	0.452	15.08	0.623
DeepLPF [8]	15.28	0.473	14.1	0.48	16.02	0.587
IPT [25]	16.27	0.504	19.8	0.813	18.3	0.811
UFormer [28]	16.36	0.771	18.82	0.771	19.66	0.871
RetinexNet [11]	16.77	0.56	15.47	0.567	17.13	0.798
Sparse [45]	17.2	0.64	20.06	0.816	22.05	0.905
EnGAN [15]	17.48	0.65	18.23	0.617	16.57	0.734
RUAS [17]	18.23	0.72	18.37	0.723	16.55	0.652
FIDE [46]	18.27	0.665	16.85	0.678	15.2	0.612
DRBN [47]	20.13	0.83	20.29	0.831	23.22	0.927
KinD [13]	20.86	0.79	14.74	0.641	13.29	0.578
Restormer [30]	22.43	0.823	19.94	0.827	21.41	0.83
MIRNet [48]	24.14	0.83	20.02	0.82	21.94	0.876
SNR-Net [26]	24.61	0.842	21.48	0.849	24.14	0.928
Retinexformer	25.16	0.845	22.8	0.84	25.67	0.93
Our method	25.92	0.87	24.34	0.86	27.81	0.96

The results demonstrate that our approach surpasses other techniques within this dataset.

In the second step, we will compare the visual effects of these methods. We processed the images from the LOL dataset using our algorithm as well as several popular algorithms, obtaining visualization results for each.

Due to space constraints, we selected the results of four images as representatives for display. We presented the original images along with the results processed using 3D-LUT, RetFormer, RUAS, Retinexformer, and our method. The results are shown in Fig. 5. Specifically, the first row contains the original images, the second row shows the images processed with 3D-LUT, the third row displays the images obtained using

RetFormer, the fourth row presents the images derived from RUAS, the fifth row illustrates the images acquired with Retinexformer, and the sixth row features the images resulting from our algorithm, and the seventh row is the ground truth.

Figure 5: These images are the visualization results obtained by processing the LOL dataset with different algorithms: The top row presents the original images, the second row shows the data processed using 3D-LUT, the third row shows the data processed using RetFormer, the fourth row shows the data obtained using RUAS, the fifth row shows the data obtained using Retinexformer, and the sixth row shows the data obtained using the algorithm proposed in this paper, and the seventh row is the ground truth

From the visualization results in Fig. 5, we can see that the original images are low-light images which lack many details. Previous methods such as 3D-LUT, RetFormer, RUAS, and Retinexformer either have color distortion issues, noise problems, or are not well-optimized in detail processing. Our method has been successful in addressing color distortion and is capable of significantly improving the visibility in poorly lit and low-contrast areas, efficiently eliminating noise without the introduction of speckles and artifacts, and consistently maintaining color integrity.

To clearly demonstrate the advantage of our method in detail processing, we have locally enlarged the results of different algorithms processing the same image from the LOL dataset. Fig. 6 shows the visualization results of 3DLT, Retinexformer, our method, and Ground Truth. Visually, we can see that our results are closer to the Ground Truth.

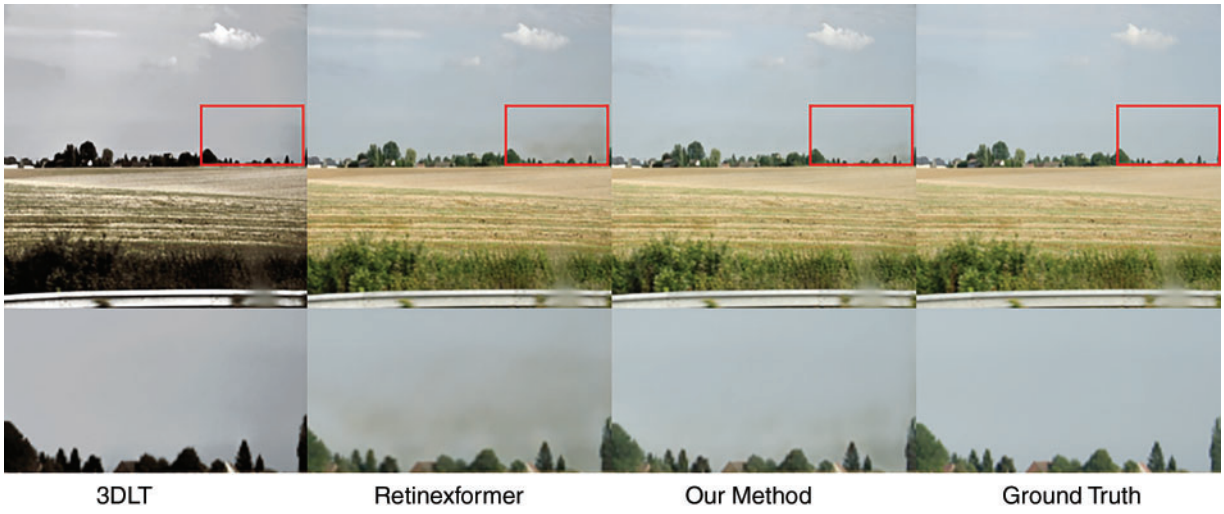

Figure 6: This figure presents the outcomes of enlarging the processing of the same image in the LOL dataset using different algorithms. The first column uses the 3DLT algorithm, the second column uses Retinexformer, the third column uses our method, and the fourth column is the ground truth

We attribute the superior visual effects of our method to two main reasons. The first part is that it's easy to obtain global information through the Transformer due to the Transformer's self-attention mechanism, which enables it to capture long-range dependencies within images. The second part is that it achieves a dual refinement of the input features because of CBAM combination of channel attention and spatial attention. This design allows the model to simultaneously focus on which channels and spatial locations are significant, thereby enhancing the model's representational power and decision-making accuracy. When combined, the two complement each other, enhancing the overall performance of the model.

To more quantitatively evaluate visual effects, we also test the LPIPS of these images. LPIPS, which stands for Learned Perceptual Image Patch Similarity, is a method for measuring the similarity between images. It uses deep learning models to assess the perceptual differences between two images. The core idea of LPIPS is to employ pre-trained deep networks to extract image features and then calculate the distance between these features to evaluate the perceptual similarity between images.

In this experiment, we select AlexNet as the pre-trained deep network and use image samples from the LOL dataset processed by 3D-LUT, RetFormer, RUAS, Retinexformer, and our own algorithm. We calculate the LPIPS values for these image samples compared to the ground truth images. The experimental results

are shown in Table 2. From the table, we can see that our method's LPIPS is significantly lower than that of other methods, indicating a clear advantage on this metric.

Table 2: LPIPS between images processed by different algorithms and the ground truth image

Methods	3D-LUT	RetFormer	RUAS	Retinexformer	Our algorithm
LPIPS	0.1566	0.0171	0.1544	0.0232	0.0096

We conduct a break-down ablation on the LOL dataset to study the effect of each component towards higher performance, as shown in Table 3. IE is derived by removing CBAM and IGAB. When we respectively apply CBAM and IGAB, IE achieves improvements significantly. This evidence suggests the effectiveness of our methods.

Table 3: Ablation studies on LOL v1

IE	CBAM	IGAB	PSNR	SSIM
✓			22.15	0.74
✓	✓		24.38	0.82
✓	✓	✓	25.92	0.87

5 Conclusion

In this paper, we introduce an approach that combines the Convolutional Block Attention Module (CBAM) with the Transformer framework to improve low-light image quality. It is based on Retinex theory. It consists of illumination estimation and a corruption restorer. The corruption is represented with a collaborative network formed by convolutional networks, CBAM, and illumination-Guided Attention Block (IGAB). CBAM only slightly increases the number of parameters, and the entire training process remains highly efficient. Experimental results demonstrate that this novel method outperforms previous approaches.

Although this algorithm has achieved better performance compared to other existing methods, two aspects deserve further attention. On one hand, it is designed based on a simple improvement of Retinex theory, but this improvement is not comprehensive and may limit the capabilities of our algorithm. On the other hand, it is obtained through supervised learning with limited datasets and advanced algorithms to produce a usable model, and the amount of training data will limit the final performance of our algorithm. Therefore, whether there will be more comprehensive principles in the future and whether semi-supervised methods can be combined for sample expansion to achieve better experimental results are both directions worth exploring in subsequent research.

Acknowledgement: We are very grateful to Chen Wei [11] and Wenhan Yang [45] for creating and sharing the LOL (Low Light Benchmark) dataset, which has provided the foundational data for the research in this paper, allowing us to carry out the related research smoothly.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Study conception and design, analysis and interpretation of results: Zhefeng Sun; references collection: Chen Wang; draft manuscript preparation: Zhefeng Sun, Chen Wang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: This research was carried out in compliance with ethical guidelines.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Chen C, Chen Q, Xu J, Koltun V. Learning to see in the dark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA; 2018. p. 3291–300.
2. Lv F, Lu F, Wu J, Lim C. MBLLEN: low-light image/video enhancement using CNNs. In: British Machine Vision Conference (BMVC); 2018 Sep 3–6; Newcastle, UK.
3. Ignatov A, Kobyshev N, Timofte R, Vanhoey K. Dslr-quality photos on mobile devices with deep convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice, Italy. p. 3297–305.
4. Lore KG, Akintayo A, Sarkar S. LLNet: a deep autoencoder approach to natural low-light image enhancement. Pattern Recognit. 2017;61:650–62. doi:10.1016/j.patcog.2016.06.008.
5. Gharbi M, Chen J, Barron JT, Hasinoff SW, Durand F. Deep bilateral learning for real-time image enhancement. ACM Trans Grap (TOG). 2017;36(4):1–12. doi:10.1145/3072959.307359.
6. Cai J, Gu S, Zhang L. Learning a deep single image contrast enhancer from multi-exposure images. IEEE Trans Image Process. 2018;27(4):2049–62. doi:10.1109/TIP.2018.2794218.
7. Guo C, Li C, Guo J, Loy CC, Hou J, Kwong S, et al. Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA; 2020. p. 1777–86.
8. Moran S, Marza P, McDonagh S, Parisot S, Slabaugh G. DeepLPF: deep local parametric filters for image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020; Seattle, WA, USA. p. 12823–32.
9. Xu X, Wang R, Lu J. Low-light image enhancement via structure modeling and guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 18–22; Vancouver, BC, Canada. p. 9893–903.
10. Wu Y, Pan C, Wang G, Yang Y, Wei J, Li C, et al. Learning semantic-aware knowledge guidance for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 18–22; Vancouver, BC, Canada. p. 1662–71.
11. Wei C, Wang W, Yang W, Liu J. Deep Retinex decomposition for low-light enhancement. In: British Machine Vision Conference; 2018 Sep 3–6; Newcastle, UK.
12. Wang W, Wei C, Yang W, Liu J. GLADNet: low-light enhancement network with global awareness. In: 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018); 2018 May 15–19; Xi'an, China: IEEE. p. 751–5.
13. Zhang Y, Zhang J, Guo X. Kindling the darkness: a practical low-light image enhancer. In: Proceedings of the 27th ACM International Conference on Multimedia; 2019 Oct 21–25; Nice, France. p. 1632–40.
14. Wang R, Zhang Q, Fu CW, Shen X, Zheng WS, Jia J. Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA, USA. p. 6849–57.
15. Jiang Y, Gong X, Liu D, Cheng Y, Fang C, Shen X, et al. EnlightenGAN: deep light enhancement without paired supervision. IEEE Trans Image Process. 2021;30:2340–9. doi:10.48550/arXiv.1906.06972.
16. Yang W, Wang S, Fang Y, Wang Y, Liu J. From fidelity to perceptual quality: a semi-supervised approach for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 3060–9.

17. Liu R, Ma L, Zhang J, Fan X, Luo Z. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 10556–65.
18. Wu W, Weng J, Zhang P, Wang X, Yang W, URetinex-Net Jiang J. URetinex-Net Retinex-based deep unfolding network for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 5891–900.
19. Fu Z, Yang Y, Tu X, Huang Y, Ding X, Ma KK. Learning a simple low-light image enhancer from paired low-light instances. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. p. 22252–61.
20. Wang Y, Wan R, Yang W, Li H, Chau LP, Kot A. Low-light image enhancement with normalizing flow. Proc AAAI Conf Artif Intell. 2022;36(3):2604–12. doi:10.1609/aaai.v36i3.20162.
21. Ye D, Ni Z, Yang W, Wang H, Wang S, Kwong S. Glow in the dark: low-light image enhancement with external memory. IEEE Trans Multimedia. 2023;26:2148–63. doi:10.1109/TMM.2023.3293736.
22. Fei B, Lyu Z, Pan L, Zhang J, Yang W, Luo T, et al. Generative diffusion prior for unified image restoration and enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 9935–46.
23. Ma L, Ma T, Liu R, Fan X, Luo Z. Toward fast, flexible, and robust low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 5627–36.
24. Zeng H, Cai J, Li L, Cao Z, Zhang L. Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time. IEEE Trans Pattern Anal Mach Intell. 2020;44(4):2058–73. doi:10.1109/TPAMI.2020.3026740.
25. Chen H, Wang Y, Guo T, Xu C, Deng Y, Liu Z, et al. Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 12294–305.
26. Xu X, Wang R, Fu CW, Jia J. SNR-aware low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 17693–703.
27. Cui Z, Li K, Gu L, Su S, Gao P, Jiang Z, et al. You only need 90k parameters to adapt light: a lightweight transformer for image enhancement and exposure correction. In: British Machine Vision Conference; 2022 Nov 21–24; London, UK.
28. Wang Z, Cun X, Bao J, Zhou W, Liu J, Li H. Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 17662–72.
29. Fu Y, Hong Y, Chen L, You S. LE-GAN: unsupervised low-light image enhancement network using attention module and identity invariant loss. Knowl-Based Syst. 2022;240:108010. doi:10.1016/j.knosys.2021.108010.
30. Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M. ResTormer: efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 5718–29.
31. Wang T, Zhang K, Shen T, Luo W, Stenger B, Lu T. Ultra-high-definition low-light image enhancement: a benchmark and transformer-based method. Proc AAAI Conf Artif Intell. 2023;37(3):2654–62.
32. Dang J, Li Z, Zhong Y, Wang L. WaveNet: wave-aware image enhancement. In: Proceedings of the Pacific Conference on Computer Graphics and Applications; 2023 Oct 10–13; Daejeon, Republic of Korea; p. 21–9.
33. Wang C, Wu H, Jin Z. FourLLIE: boosting low-light image enhancement by fourier frequency information. In: Proceedings of the 31st ACM International Conference on Multimedia; 2023 Oct 29–Nov 3; Ottawa, ON, Canada. p. 7459–69.
34. Wang Y, Liu Z, Liu J, Xu S, Liu S. Low-light image enhancement with illumination-aware gamma correction and complete image modelling network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 1–6; Paris, France. p. 13082–91.

35. Liu Y, Huang T, Dong W, Wu F, Li X, Shi G. Low-light image enhancement with multi-stage residue quantization and brightness-aware attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 1–6; Paris, France. p. 12140–49.
36. Yi X, Xu H, Zhang H, Tang L, Ma J. Diff-Retinex: rethinking low-light image enhancement with A generative diffusion model. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France: IEEE. p. 12268–77.
37. Dang J, Zhong Y, Qin X. PPformer: using pixel-wise and patch-wise cross-attention for low-light image enhancement. Comput Vis Image Underst. 2024;241:103930.
38. Cai Y, Bian H, Lin J, Wang H, Timofte R, Zhang Y. Retinexformer: one-stage Retinex-based transformer for low-light image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 1–6; Paris, France. p. 12504–13.
39. Xu K, Ba JL, Kiros R, Cho K, Courville A, Salakhutdinov R, et al. Attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning; 2015 Jul 6–11; Lille, France. p. 2048–57.
40. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst (NIPS). In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS); 2017 Dec 4–9; Long Beach, CA, USA. 2017. p. 5998–6008.
41. Zhou G, Zhu X, Song C, Fan Y, Zhu H, Ma X, et al. Deep interest network for click-through rate prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018 Aug 19–23; London, UK. p. 1059–68.
42. Woo S, Park J, Lee JY, Kweon IS. Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany. p. 3–19.
43. Fu X, Zeng D, Huang Y, Zhang XP, Ding X. A weighted variational model for simultaneous reflectance and illumination estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 2782–90.
44. Guo X, Li Y, Lime Ling H. LIME: Low-light image enhancement via illumination map estimation. IEEE Trans Image Process. 2016;26(2):982–93.
45. Yang W, Wang W, Huang H, Wang S, Liu J. Sparse gradient regularized deep Retinex network for robust low-light image enhancement. IEEE Trans Image Process. 2021;30:2072–86. doi:10.1109/TIP.2021.3050850.
46. Xu K, Yang X, Yin B, Lau RWH. Learning to restore low-light images via decomposition-and-enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 2278–87.
47. Yang W, Wang S, Fang Y, Wang Y, Liu J. Band representation-based semi-supervised low-light image enhancement: bridging the gap between signal fidelity and perceptual quality. IEEE Trans Image Process. 2021;30:3461–73. doi:10.1109/TIP.2021.3062184.
48. Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang MH, et al. Learning enriched features for real image restoration and enhancement. In: Proceedings of the European Conference on Computer Vision (ECCV); 2020 Aug 23–28. p. 492–511.