Mathematical Biosciences and Engineering

MBE, 21(2): 1917–1937.
DOI: 10.3934/mbe.2024085
Received: 25 September 2023
Revised: 20 December 2023
Accepted: 25 December 2023
Published: 05 January 2024

http://www.aimspress.com/journal/MBE

*Research article*

# HPCDNet: Hybrid position coding and dual-frquency domain transform network for low-light image enhancement

**Mingju Chen[1,2], Hongyang Li[1,*], Hongming Peng[1], Xingzhong Xiong[1,2] and Ning Long[3]**

[1] School of Automation and Information Engineering, Sichuan University of Science & Engineering, Yibin 644002, China
[2] Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science & Engineering, Yibin 644002, China
[3] School of Network & Communication Engineering, Chengdu Technological University, Chengdu 611730, China

* **Correspondence:** Email: 322085404523@stu.suse.edu.cn.

**Abstract:** Low-light image enhancement (LLIE) improves lighting to obtain natural normal-light images from images captured under poor illumination. However, existing LLIE methods do not effectively utilize positional and frequency domain image information. To address this limitation, we proposed an end-to-end low-light image enhancement network called HPCDNet. HPCDNet uniquely integrates a hybrid positional coding technique into the self-attention mechanism by appending hybrid positional codes to the query and key, which better retains spatial positional information in the image. The hybrid positional coding can adaptively emphasize important local structures to improve modeling of spatial dependencies within low-light images. Meanwhile, frequency domain image information lost under low-light is recovered via discrete wavelet and cosine transforms. The resulting two frequency domain feature types are weighted and merged using a dual-attention module. More effective use of frequency domain information enhances the network's ability to recreate details, improving visual quality of enhanced low-light images. Experiments demonstrated that our approach can heighten visibility, contrast and color properties of low-light images while better preserving details and textures than previous techniques.

**Keywords:** image enhancement; position coding; wavelet transform; cosine transform; self-attention

## 1. Introduction

Computer vision encompasses various pivotal tasks and methods like object detection, image classification, segmentation, deblurring and facial recognition [1–3]. However, these techniques presume well-lit input images, whereas real-world data is often dim and degraded. Such low-light images confound both human and machine perception, undermining subsequent analysis and impeding real-world computer vision applications. Addressing low-light images is thus imperative for computer vision. Low-light enhancement techniques aim to elucidate obscured content and avert performance declines in downstream tasks. Early methods like histogram equalization [4] and Retinex [5] have limited efficacy and generalizability for low-light enhancement.

In recent years, numerous deep learning-based low-light image enhancement (LLIE) techniques have emerged. The advantage of deep learning is its ability to learn complex feature representations from large amounts of data and optimize them through a training process leading to more accurate image enhancement. These methods fall into two categories: end-to-end framework [6–13] and those based on Retinex [14–21]. Among them, Jiang et al. [22] proposed an unsupervised generative adversarial network framework for enhancement of low-light images that does not require paired images for training. Zhang et al. [23] proposed an unsupervised learning method for enhancement of low-light images, which utilizes the prior knowledge of histogram equalization to guide the network in learning the enhancement mappings. Deep learning methods enhance low-light images by modeling the relationship between low-light and high-quality images, generally outperforming traditional techniques.

However, many existing methods ignore the position information of the image, which can lead to a lack of spatial coherence and difficulty in preserving the detailed texture of the image. In addition, there have been researches proving the importance of position information to the image. Dosovitskiy et al. [24] achieved state-of-the-art results in large-scale image classification tasks by using sine-based absolute position coding in Vision Transformer. In 3D target detection, Xu et al. [25] achieved a more effective aggregation of the two feature types by adding position coding between voxel features and original point features, thus improving the detection accuracy. All these methods demonstrate the importance of positional information for image, and similarly, positional information is also important for LLIE.

Most LLIE techniques only consider spatial domain information, neglecting the usefulness of frequency domain cues for quality improvement. The frequency domain harbors data regarding an image's frequency components, offering salient insights into texture, detail, and structure. To better leverage frequency domain data, some studies have incorporated techniques including wavelet and cosine transforms into their networks. Wavelet transforms can decompose low-frequency illumination and high-frequency details, aiding detail and texture restoration. Cosine transforms can concentrate image information in lower frequencies, enhancing brightness and contrast. For example, Fan et al. [7] enriched the wavelet domain features by half-wavelet attention blocks, which effectively improved the quality of the image, and Tiwari et al. [26] proposed a method to control the enhancement degree based on the cosine transform, which verified the value of the frequency domain information to a certain extent. However, these methods do not fully utilize the frequency domain information, and there is still some room for improvement, especially in lighting and detailed texture.

To fully leverage positional and frequency domain image information, inspired by previous work [27–32], we propose an LLIE network called HPCDNet. HPCDNet more efficiently acquires

positional information of images by introducing hybrid positional encodings into the self-attention mechanism. Unlike conventional self-attention mechanisms that solely rely on absolute distances, this hybrid positional encoding considers both relative positional relationships within the image and global absolute coordinate information. Specifically, we add a unique sinusoidal encoding for each pixel position as its absolute coordinates, while also learning the relative offset relationships between each position and its surrounding pixels. Dual-frequency attention block modules are then utilized to extract frequency domain features lost under low-light conditions, which are then weighted by dual attention units (DAU) [33]. Finally, the features are converted back to the spatial domain via inverse transforms. Overall, the primary contributions of this work can be summarized as follows:

- We propose a hybrid position coding scheme for self-attention mechanisms to better capture global structure and local details in images. The hybrid scheme combines both relative and absolute position encodings to capture such multi-scale information.
- We propose an efficient feature extraction building block named dual-frequency attention block (DFAB), which extracts frequency domain features via discrete cosine transform and discrete wavelet transform and weighs these features using a DAU. By operating on both spatial and frequency domains, DFAB improves feature utilization and representation power.
- To consolidate multilevel representations, we design a cross-layer fusion block (CFB) module based on partial convolutions for adaptive integration of hierarchical features via learned cross-scale interactions.
- We propose a generalizable LLIE network called HPCDNet.We evaluate HPCDNet on the LOL and MIT-Adobe FiveK datasets. Experiments show HPCDNet significantly outperforms prior arts in low-light enhancement.

## 2. Related work

### *2.1. Low-light image enhancement*

**Traditional methods.** Traditional methods for enhancing low-light images primarily includes histogram equalization-based approaches [4] and methods based on the Retinex theory [5]. One category of methods seeks to enhance low-light images by remapping brightness levels to expand the dynamic range, which accentuates darker regions to increase visibility and quality. In contrast, Retinex-based approaches decompose images into reflectance and illumination layers, using the estimated reflectance as the final enhanced output. Thus, Retinex methods strongly rely on accurate modeling of image components and fitting prior knowledge. However, designing prior knowledge that is applicable to diverse scenarios is a challenging task [19]. To mitigate inherent constraints with conventional approaches, deep learning has been overwhelmingly embraced as a promising paradigm for low-light enhancement methodologies.

**Deep learning-based methods.** In recent years, LLIE methods based on deep learning have shown good results. Lore et al. [34] proposed a method based on deep autoencoders to adaptively enhance images by identifying signal features in low-light images. Wei et al. [14] proposed a deep learning-based Retinex decomposition method, which mainly enhances low-light images by learning data-driven reflection maps and illumination maps. Zhang et al. [35,36] proposed two improved methods based on Retinex-Net, called KinD and KinD++, which adjusts image brightness by introducing global and local gains. Zamir et al. [33] enhanced the details and texture of low-light

images by learning multi-scale features and combining contextual information at different scales. Recently, Wu et al. [19] also proposed a Retinex-based network that decomposes low-light images into reflection maps and illumination maps and effectively suppresses noise in the image by learning to adaptively fit implicit prior knowledge. In addition, Fan et al. [7] also proposed an image enhancement network that effectively utilized the frequency domain information of the image by using semi-wavelet attention blocks to enrich wavelet domain features. In the same year, Zhang et al. [37] proposed a new deep color consistent network termed DCC-Net, which can jointly preserve color information and enhance the illumination. Wang et al. [30] proposed a low-light enhancement method based on transformer, and improved image quality through an axis-based multi-head self-attention mechanism and a cross-layer attention fusion block. However, these methods do not fully utilize the frequency domain information of the image. In contrast, the network we designed learns the frequency domain information of images in an end-to-end manner through an innovative dual frequency domain transformation module. In this way, the network can better perceive the overall pixel distribution of the image while restoring finer and natural local details.

### *2.2. Position encoding*

In recent years, several positional coding methods have been proposed, and these methods can be categorized into two groups: Relative positional coding and absolute positional coding.

**Relative positional coding.** Shaw et al. [29] introduced a relative positional encoding approach tailored for self-attentional architectures. The input tokens are modeled as directed, fully connected graphs, and each edge between two arbitrary positions $i$ and $j$ is represented by a learnable vector. Dai et al. [38] introduced an additional bias term for the query and use a sinusoidal function for relative position encoding. Recently, Huang et al. [39] proposed a new approach that simultaneously considers the interaction of query, key and relative positions. While Ramachandran et al. [40] proposed a position coding method for images, which divides the 2D relative coding into horizontal and vertical directions so that each direction can be modeled by a 1D coding, Wang et al. [41] introduced a position-sensitive method by incorporating a qkv-dependent positional bias in the self-attention. Inspired by their predecessors, Wu et al. [42] proposed a new relative position coding method specifically for images, called image RPE (iRPE), which takes into account the modeling of relative position distances in a self-attention mechanism and the interaction between query and relative position embedding.

**Absolute positional coding.** Parmar et al. [43] proposed an image transformer framework, using transformer in image tasks for the first time, and by using absolute position encoding to introduce spatial information, they proved its effectiveness in tasks such as image classification and segmentation. Dosovitskiy et al. [24] used absolute position coding in the ViT model and achieved good results on image classification tasks, which reflects the importance of absolute position coding in image classification tasks. In the same year, Carion et al. [44] proposed the first end-to-end transformer target detection model DETR, which also used absolute position encoding to obtain spatial prior knowledge. In addition, Xie et al. [45] designed a SegFormer model suitable for image segmentation, which also used absolute position coding to provide spatial information to the transformer, and achieved optimal segmentation performance on multiple datasets.

Notwithstanding the achievements of prevailing techniques, the value of positional encoding remains underestimated. Current methods generally adopt either relative or absolute positional coding exclusively; thus, they are unable to fully leverage the spatial knowledge encompassed within the

image. Discerning that the local orientation cues of relative positional coding and global coordinate information of absolute positional coding are complementary, we consolidate both schemes for LLIE to empower superior modeling of the spatial structure.

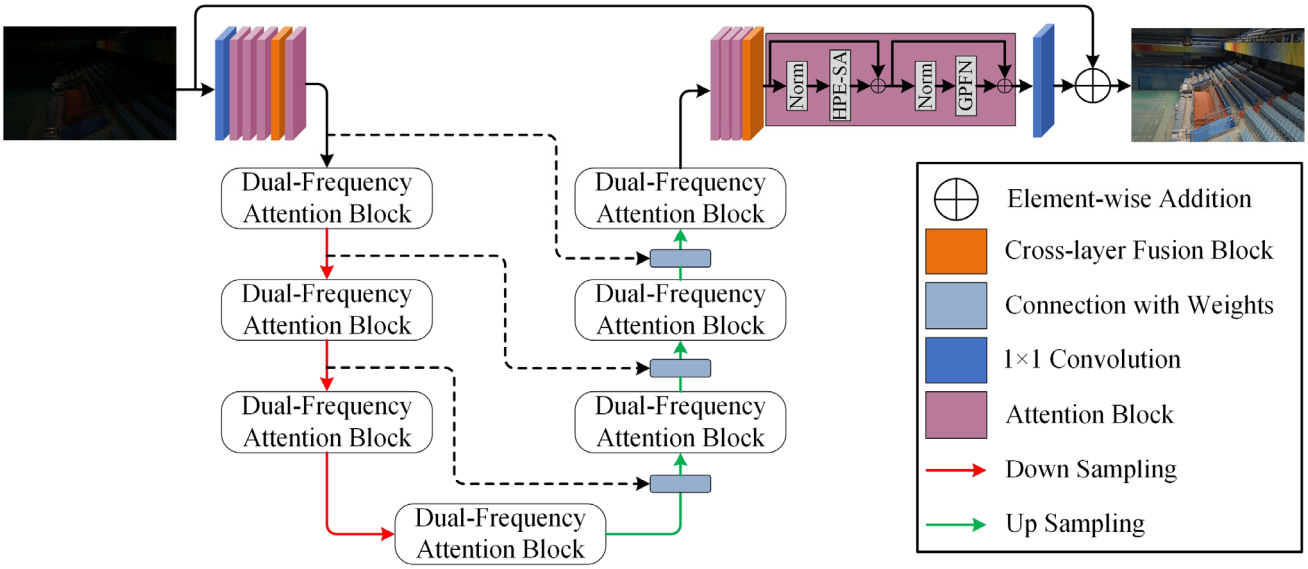


**Figure 1.** The overall structure of our proposed HPCDNet. Due to our unique design, it can better capture the position and frequency domain information of the image.

## 3. Proposed method

This section delineates the proposed LLIE network architecture depicted in Figure 1. The framework incorporates four pivotal constituents: (a) Hybrid position encoding self-attention (HPE-SA): This module is used to perform nonlinear transformation and combination of input image features to improve the performance of the model. Attention mechanism plays a crucial role in modeling and capturing global contextual information in images, and the self-attention mechanism with hybrid positional encoding can better preserve the image spatial structure. (b) Gated-pconv feed-forward network (GPFN): Employs partial convolutions alongside nonlinearities to learn representations for enriching details and enhancing visual quality. (c) DFAB: Leverages frequency domain knowledge to adaptively aggregate input characteristics via attention-based merging to heighten global detail while preserving textures. (d) CFB: Aggregates multi-scale representations by weighting crucial characteristics, thereby upholding holistic structure alongside enriching local cues.

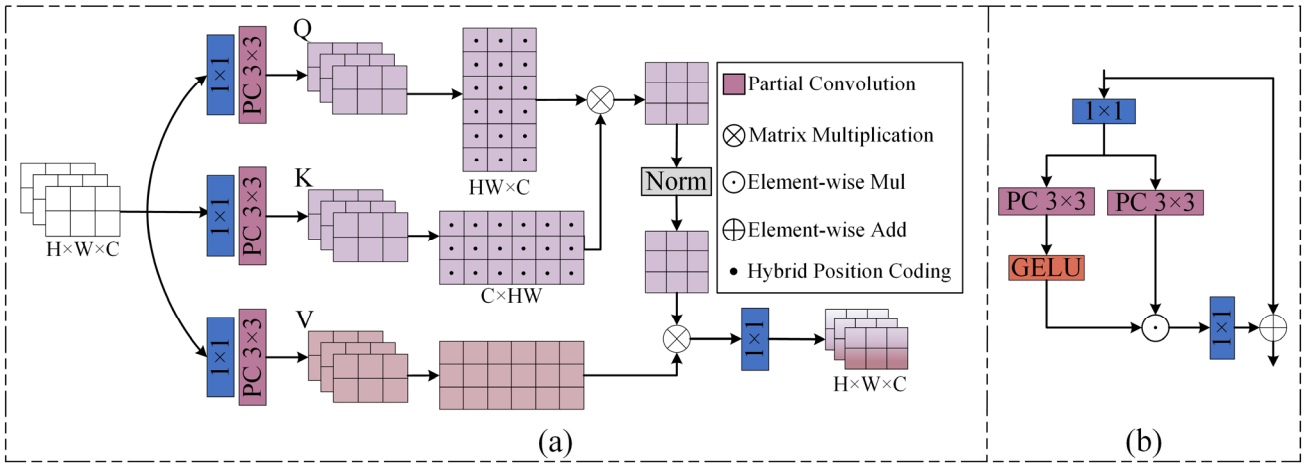


**Figure 2.** Main components in attention block, from left to right in order (a) HPE-SA and (b) GPFN.

### *3.1. Hybrid position encoding self-attention*

For LLIE, positional cues are especially pivotal for modeling spatial correlation amongst pixels. Specifically, lacking such localization knowledge hinders comprehensive characterization of inter-region connections required for effective image upgrading. Moreover, in the traditional self-attention mechanism, the time and memory complexity of the key-query dot-product interaction grows quadratically with the spatial resolution of input, i.e., $\mathcal{O}(W^2H^2)$ for images of $W \times H$ pixels, which means that when applying the self-attention mechanism in LLIE tasks, you may encounter the problem of insufficient computing resources.

To alleviate these problems, inspired by [3,27,28,30,32], we propose HPE-SA, a self-attention mechanism with hybrid position encoding, whose structure is shown in Figure 2(a). The HPE-SA module greatly reduces the amount of computation by performing self-attention calculation across the feature dimension instead of the spatial dimension, and by specifically fusing local and global contextual information before doing so, it applies the attention mechanism to $C$ feature channels instead of to $HW$ spatial locations. It computes a transposed attention map of size $C \times C$, so the complexity is $\mathcal{O}(C^2)$; $C$ is usually much smaller than $HW$ and the number of feature channels is much smaller than the number of pixels, so $\mathcal{O}(C^2)$ can be regarded as a constant level complexity. Even if the image size increases, $C$ does not change, so the complexity is not affected by $H$, $W$. Therefore, it can be said that there is a linear relationship between the computational complexity of the HPE-SA module and the spatial size ($H$, $W$). This substantially alleviates complexity versus conventional self-attention models. Additionally, hybrid positional encodings empower the attention mechanism to better model nuanced inter-pixel relationships, simultaneously refining local details as well as global contexts.

In detail, for an input tensor of size $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, we first obtain a representation with local context information through a 1 × 1 convolution. Following this, we encode the spatial context using a 3 × 3 partial convolution, resulting in $\mathbf{Q} = W^{Q}_{P3\times3}W^{Q}_{1\times1}\mathbf{X}$, $\mathbf{K} = W^{K}_{P3\times3}W^{K}_{1\times1}\mathbf{X}$ and $\mathbf{V} = W^{V}_{P3\times3}W^{V}_{1\times1}\mathbf{X}$. Here, $W^{(\cdot)}_{P3\times3}$ and $W^{(\cdot)}_{1\times1}$ denote 3 × 3 partial convolution and 1 × 1 ordinary convolution, respectively. Thus, it can be expressed as:

$$\widehat{\mathbf{X}} = \widehat{\mathbf{V}}\,\mathrm{Softmax}\left(\widehat{\mathbf{Q}}\cdot\widehat{\mathbf{K}}\right), \tag{1}$$

Subsequently, hybrid positional encodings are formulated to augment queries and keys with both absolute spatial coordinates alongside relative pixel displacements. Specifically, this hybrid representation comprises distinct absolute and relative positional encodings, detailed as follows.

Absolute position coding can provide additional spatial information to help the attention mechanism better model the dependencies between different positions of an image. We use an absolute position coding method based on sine and cosine; specifically, a position coding matrix $P_{abs}$ can be constructed, and the formula for $P_{abs}$ is as follows:

$$P_{abs}(i,j,k) = \begin{cases} sin(k\dfrac{2i\pi}{C}e^{-\frac{ln(1000)}{C/2}i}), & i\ is\ even \\ cos(k\dfrac{2i\pi}{C}e^{-\frac{ln(1000)}{C/2}i}), & i\ is\ odd \\ sin(k\dfrac{2j\pi}{C}e^{-\frac{ln(1000)}{C/2}j}), & j\ is\ even \\ cos(k\dfrac{2j\pi}{C}e^{-\frac{ln(1000)}{C/2}j}), & j\ is\ odd \end{cases}, \tag{2}$$

where $P_{abs} \in \mathbb{R}^{H\times W\times C}$ denotes the absolute position encoding matrix, *i* and *j* denote the row and column indices in the absolute position encoding matrix, *k* denotes the index of the third dimension of the $P_{abs}$ matrix, and *d* denotes the dimension of the model. We then decompose $P_{abs}$ into $P_{abs}^{Q}$ and $P_{abs}^{K}$ to obtain the absolute position encoding matrices of $\widehat{\mathbf{Q}}$ and $\widehat{\mathbf{K}}$.

While integration of absolute positional encoding imparts global localization, relative spatial offsets across different areas remain equally vital. Hence, we supplement relative positional encoding to facilitate improved modeling of inter-pixel spatial correlations. Nevertheless, existing relative offset encodings condition solely on the query without considering the key. Since key could indicate attention-worthy regions, we propose a trainable relative encoding mechanism. Specifically, we first construct a trainable encoding matrix $E \in \mathbb{R}^{C\times(2D-1)}$, where *C* is the total number of channels of Q and K, and $2D-1$ represents the range of relative positions in the two-dimensional space. Each row vector of matrix *E* encodes the relative displacement between any two positions. Next, we precalculate a relative position index matrix $M_{ij} \in \mathbb{R}^{D\times D}$ based on the row and column coordinates, where $M(i,j) = i - j + D - 1$. For efficient indexing, we flatten *M* into a one-dimensional index vector $I \in \mathbb{R}^{D^2}$, then retrieve the corresponding encoding vector $P = E[:, I] \in \mathbb{R}^{C\times D^2}$ from the encoding matrix *E* based on the index vector *I*. Finally, we split the coding tensor *P* into Q and K in proportion to the number of channels *C*, so that we can obtain $PE_r^Q$ and $PE_r^K$, then we can obtain the positional bias $\widehat{\mathbf{Q}}\cdot PE_r^Q$ associated with the query and the positional bias term $\widehat{\mathbf{K}}\cdot PE_r^K$ associated with the key pixel. In this way, combined with the absolute position encoding proposed above, the specific representation of HPE-SA can be expressed as follows:

$$\widehat{\mathbf{X}} = W_p\,\widehat{\mathbf{V}}\cdot \mathrm{LN}\left(\mathrm{Softmax}\left((\widehat{\mathbf{Q}}\cdot\widehat{\mathbf{K}} + \widehat{\mathbf{Q}}\cdot P^{Q} + \widehat{\mathbf{K}}\cdot P^{K})/\alpha\right)\right), \tag{3}$$

$$P^{Q} = P^{Q}_{rel} + P^{Q}_{abs} \tag{4}$$

$$P^{K} = P^{K}_{rel} + P^{K}_{abs} \tag{5}$$

where $\widehat{\mathbf{X}}$ represents the output feature map, $\widehat{\mathbf{Q}} \in \mathbb{R}^{HW \times C}$, $\widehat{\mathbf{K}} \in \mathbb{R}^{C \times HW}$ and $\widehat{\mathbf{V}} \in \mathbb{R}^{C \times HW}$, $PE^{Q}$ and $PE^{K}$, respectively denote the hybrid position encoding for $\widehat{\mathbf{Q}}$ and $\widehat{\mathbf{K}}$, LN stands for layer normalization [44] and $\alpha$ is a learnable scaling parameter used to regulate the magnitude of the dot product of $\widehat{\mathbf{Q}}$ and $\widehat{\mathbf{K}}$ before it is input into the softmax function.

### *3.2. Gated-pconv feed-forward network*

Conventional feedforward networks are often limited in modeling intrinsic local cues embedded within images. To address such limitations, we conceive a new feedforward formulation termed GPFN toward improved characterization of localized features. The structure is illustrated in Figure 2(b), and this network consists of two parallel branches. The former enacts 3 × 3 partial convolutions with GELU [46] activation function to extract structural impressions, while the latter aggregates contextual inter-pixel knowledge via a 3 × 3 partial convolutional layer. Subsequently, we perform element-wise multiplication between the output feature map of the second branch and the feature map of the first branch. This operation enhances the representation of crucial features while mitigating the influence of less important ones. With this innovative network design, the model becomes more adept at learning and representing local features within the image. Given an input tensor $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, the formula for GPFN is as follows:

$$\widehat{\mathbf{X}} = W_{1\times1}\,\mathrm{Gate}\left(\mathbf{X}\right) + \mathbf{X}, \tag{6}$$

$$\mathrm{Gate}\left(\mathbf{X}\right) = W^{1}_{P3\times3} W_{1\times1}(\mathbf{X}) \odot \phi(W^{2}_{P3\times3} W_{1\times1}(\mathbf{X})), \tag{7}$$

where $\odot$ represents multiplication, and $\phi$ signifies the GELU activation function. In summary, GPFN efficiently manages the flow of information across various hierarchical levels. This enables each level to focus on specific details and complement one another, ultimately enhancing the model's overall performance.

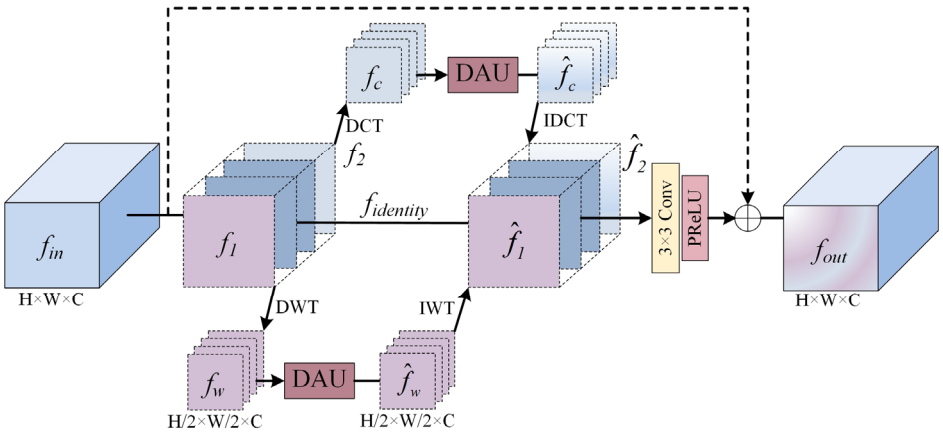


**Figure 3.** The architecture of DFAB.

### *3.3. Dual-frequency attention block*

Figure 3 illustrates the structure of the proposed DFAB, which comprises two branches corresponding to discrete wavelet transform (DWT) and discrete cosine transform (DCT). Each branch incorporates an attention mechanism block known as DAU [33,47], which includes channel attention [48] and spatial attention [49,50] to weigh the transformed feature maps. The outputs from these two branches are then fused to obtain the final feature representation. The design concept behind DFAB is to enhance the feature representation by leveraging DWT [51] and DCT [26,52]. By encoding multi-scale, multi-frequency cues, these transformations prove effective for representing structural and textural content. Within each branch, channel attention weighs the relevance of different feature channels to enable better feature selection and prioritization. Similarly, spatial attention assesses the importance of different spatial positions, allowing greater emphasis on salient areas.

The input feature $f_{in} \in \mathbb{R}^{H\times W\times C}$ is partitioned along the input channel direction into three parts: $f_1$, $f_2$ and $f_{identity}$, where $f_1 \in \mathbb{R}^{H\times W\times C/4}$, $f_1 \in \mathbb{R}^{H\times W\times C/4}$, and $f_{identity} \in \mathbb{R}^{H\times W\times C/2}$. The main purpose of dividing the input features is to reduce computational complexity and retain contextual information. Among them, $f_{identity}$ is used to preserve the normal domain features, $f_2$ performs discrete cosine transform, while $f_1$ undergoes DWT to obtain wavelet domain features $f_W$. Through DWT, the input feature is decomposed into multiple subbands, where each subband represents a different frequency range. Since these subbands contain information about different aspects of the original feature, spatial and frequency information can be better captured. These subbands are combined into a wavelet domain feature map $f_W \in \mathbb{R}^{H/2\times W/2\times C}$.

The wavelet domain feature map $f_W$ will be obtained by the DAU module with weighted wavelet domain features $\hat{f}_w \in \mathbb{R}^{H/2\times W/2\times C}$. Finally, we perform an inverse wavelet transform on the weighted wavelet domain feature $\hat{f}_w$ and reshape it to the same shape as $f_1$ and become the weighted normal domain feature $\hat{f}_1 \in \mathbb{R}^{H\times W\times C/4}$.

The other branch $f_2$ is obtained by DCT to get the feature $f_c \in \mathbb{R}^{H/2\times W/2\times C}$, and similarly, we input $f_c$ into the DAU module to obtain the weighted low-frequency feature $\hat{f}_c \in \mathbb{R}^{H/2\times W/2\times C}$. In this way, we can better utilize the low-frequency information of the input image and fuse it with the features in other domains to improve the performance of the model. We then perform an inverse cosine transform on the weighted feature $\hat{f}_c$ and reshape it to the same shape as $f_2$, which becomes the weighted normal domain feature $\hat{f}_2 \in \mathbb{R}^{H\times W\times C/4}$. These three features ($\hat{f}_1$、$\hat{f}_2$ and $f_{identity}$) are then stitched together and passed into a 3 × 3 convolutional and PReLU layer to obtain the residual features. Finally, we add the input features to the residual features to get the output features $f_{out} \in \mathbb{R}^{H\times W\times C}$.

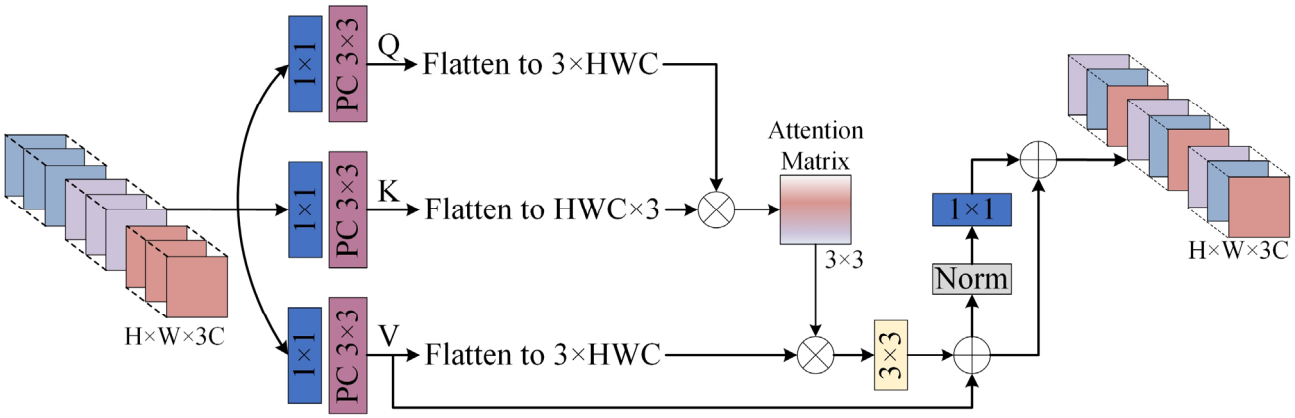


**Figure 4.** The architecture of CFB. This technique utilizes features extracted from multiple convolutional streams and then aggregates them using self-attention.

### *3.4. Cross-layer fusion block*

Recently, Zamir and Wang et al. [32,53] adopted feature joins or jump joins to aggregate representations across layers in Transformer networks. However, these operations do not fully utilize the dependencies between different layers. The low-light image contains many black zero-valued pixels, and partial convolution can effectively avoid the contamination of the results by zero-valued pixels in the convolution operation. The partial convolution only updates the valid non-zero pixels, which avoids the invalid spatial information from interfering with the results, so we designed a CFB module built upon partial convolutions, with the architecture illustrated in Figure 4. Same as self-attention in HPE-SA, we use 1 × 1 convolution to aggregate context information, and then use 3 × 3 partial convolution to generate **Q**, **K** and **V** and flatten them into a matrix of 3 × HWC. Next, we compute an attention matrix of size 3 × 3 and compute the attention weights based on the matrix and add the features of the upper **V**. Finally, we get the fused features and the process can be represented as:

$$\mathbf{F}_{\mathbf{A}} = W_{3\times3}\hat{\mathbf{V}}\,\mathrm{Softmax}(\hat{\mathbf{Q}}\cdot\hat{\mathbf{K}}/\alpha)+\hat{\mathbf{V}}, \tag{8}$$

$$\mathbf{F}_{\mathbf{out}} = W_{1\times1}\,\mathrm{LN}\left(\mathbf{F}_{\mathbf{A}}\right)+\mathbf{F}_{\mathbf{A}}, \tag{9}$$

where $\mathbf{F}_{\mathbf{out}}$ represents the output features of layers within the network that contains substantial information. We strategically incorporate CFB modules at the start and end of our pipeline. This bidirectional design enables consolidation of multilevel representations, yielding more holistic and descriptive feature embeddings.

### *3.5. Loss function*

We used a loss function in our experiments consisting of three parts, each of which is specified below:

**Smooth L1 Loss.** To encourage accurate regression for low-light enhancement and to suppress noise interference, we used the smooth L1 loss function $\mathcal{L}_{smooth-L1}$ between the predicted enhanced

image $I_e$ and the ground truth $I_{gt}$:

$$\mathcal{L}_{smooth-L1}(I_e, I_{gt}) = \begin{cases} 0.5(I_e - I_{gt})^2 & if \ |I_e - I_{gt}| < 1 \\ |I_e - I_{gt}| - 0.5, & \text{otherwise} \end{cases}, \tag{10}$$

**SSIM Loss.** Considering that the degradation of low-light images is caused by a variety of factors, in order to comprehensively evaluate the differences between images in terms of luminance, contrast, and structure, we adopt the structural similarity (SSIM) index as the loss function. Specifically, the SSIM loss $\mathcal{L}_s$ is defined as:

$$\mathcal{L}_s = 1 - \text{SSIM}(I_e, I_{gt}), \tag{11}$$

where $\text{SSIM}(\cdot)$ calculates the structural similarity between two images based on statistical measures. By minimizing $\mathcal{L}_s$, it allows the model to generate enhancement results with better perceptual consistency with the high quality reference image in terms of luminance, gradient and structural patterns.

**Perceptual Loss.** In order to utilize the semantic information to improve the visual quality of enhanced images, we adopt the perceptual loss $\mathcal{L}_p$ as the perceptual metric. Specifically, it is the Euclidean distance between the feature representations extracted from the pretrained convolutional neural network. Given the enhanced image $I_e$ and the ground truth $I_{gt}$, the mathematical definition of the perceptual loss $\mathcal{L}_p$ is as follows:

$$\mathcal{L}_p = \frac{1}{WHC} \| \phi(I_{gt}) - \phi(I_e) \|^2, \tag{12}$$

where $W$, $H$ and $C$ denote the three dimensions of the image, respectively, and $\phi(\cdot)$ denotes the pre-trained VGG network [54].

**Total Loss.** By combining $\mathcal{L}_{smooth-L1}$, $\mathcal{L}_s$ and $\mathcal{L}_p$, we can get the total loss function $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{smooth-L1} + \mathcal{L}_s + \mathcal{L}_p, \tag{13}$$

## 4. Experimental section

In this section, we present the implementation details of our experiments. We offer visualizations to showcase the comparison between the images generated by our model and those produced by other algorithms. We then evaluate and compare the results generated by our model with those of previous methods. First, we conduct a quantitative evaluation using a range of commonly employed image quality assessment metrics, including PSNR, SSIM and LPIPS. These metrics help measure the similarity between the images generated by our model and the ground truth. Second, these visual comparisons provide insights into the qualitative performance of our approach in enhancing low-light images.

### *4.1. Experimental details*

The experiments in this paper were conducted using an NVIDIA GTX TITAN Xp GPU and

PyTorch 2.0.0. The network was trained exclusively on images with dimensions of 128 × 128, utilizing a batch size of four for a total of 1200 iterations. The Adam optimizer was employed, with an initial learning rate set to $1 \times 10^{-4}$. The learning rate was then reduced to $1 \times 10^{-6}$ following the cosine decay strategy. Finally, the model was tested on the LOL dataset and images from the MIT-Adobe FiveK dataset.

### *4.2. Experiments and result*

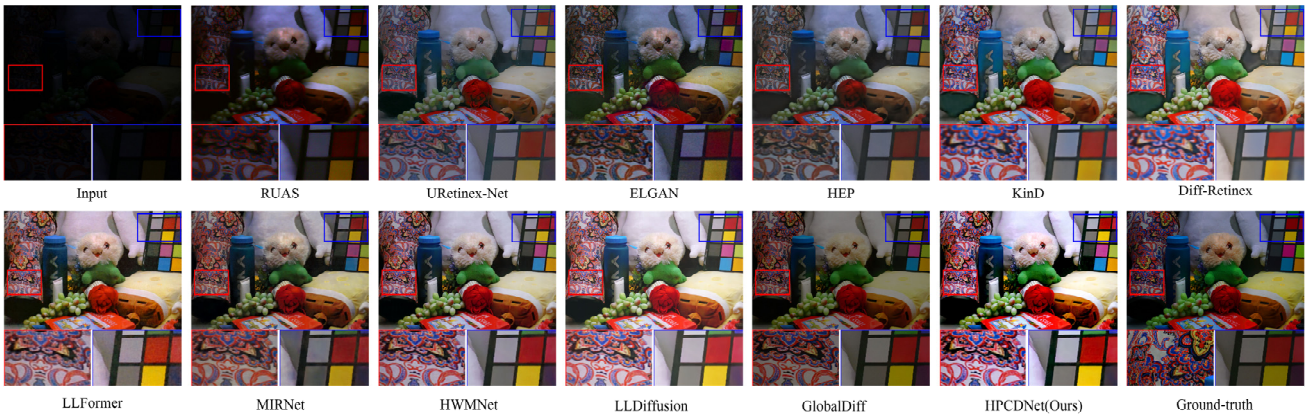


**Figure 5.** Visual comparison on the LoL dataset [14]. We performed a visual comparison of low-light enhancement methods. Our method outperforms other methods in enhancing images with complex colors and textures.

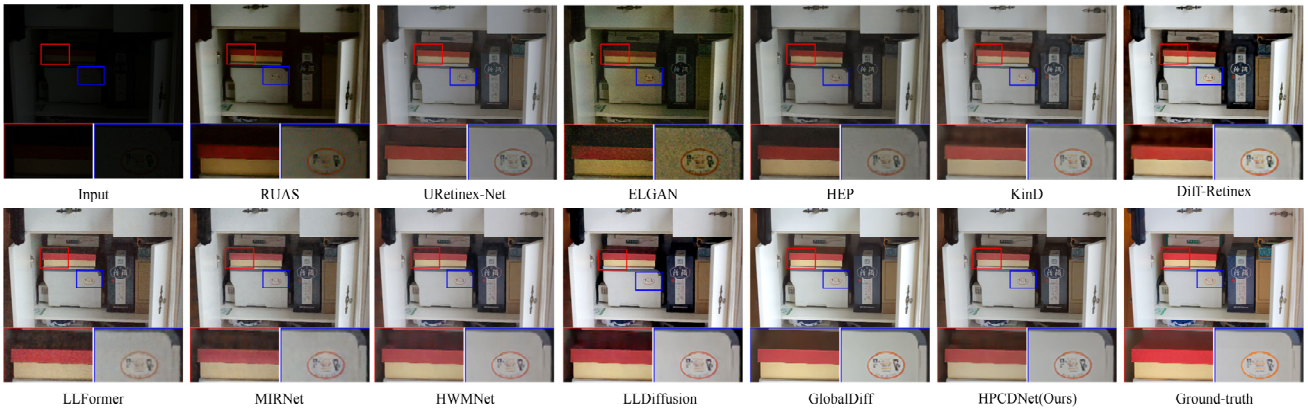


**Figure 6.** Visual comparison on the LoL dataset [14]. We performed a visual comparison of low-light enhancement methods.

Figures 5 and 6 present qualitative comparisons on the LOL dataset against several state of the art methods. Specifically, we compare against leading existing techniques. It can be observed that KinD, LLFormer and MIRNet introduce noticable noise, especially for images with vibrant colors and intricate textures. In contrast, our proposed approach demonstrates slight improvements over LLDiffusion in reconstructing such challenging cases. Additionally, our technique showcases strengths in preserving textural details while heightening contrast. Notably, our method restores more naturalistic color patterns resembling real-world scenarios.

**Table 1.** Comparison of low-light enhancement methods performed on the LoL dataset [14]; ↑ (↓) denotes that larger (smaller) values lead to better performance; the times in the table represent the total time taken by the model to reason about the 15 images on the test set.

| Method | PSNR↑ | SSIM↑ | LPIP↓ | Params | Times | Method | PSNR↑ | SSIM↑ | LPIP↓ | Params | Times |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RUAS [17] | 16.41 | 0.50 | 0.27 | 0.003 M | 0.02s | LLFormer [30] | 23.65 | 0.82 | 0.17 | 24.5M | 0.34 s |
| Uretinex [19] | 16.77 | 0.56 | 0.47 | 6.7 M | 0.37s | MIRNet [33] | 24.14 | 0.83 | 0.13 | 31.8M | 0.16 s |
| ELGAN [22] | 17.48 | 0.65 | 0.32 | 7.0 M | 0.18s | HWMNet [7] | 24.24 | 0.85 | 0.12 | 3.6M | 0.47 s |
| HEP [23] | 20.23 | 0.79 | 0.19 | 2.9 M | 0.06s | LLDiffusion [55] | 24.65 | 0.85 | 0.08 | — | — |
| KinD [36] | 20.87 | 0.80 | 0.17 | 8.5 M | 0.02s | GlobalDiff [56] | 27.83 | 0.87 | 0.09 | 17.4M | 0.93 s |
| Diff-Retinex [57] | 21.98 | 0.86 | 0.05 | — | — | **HPCDNet(Ours)** | **24.83** | **0.87** | **0.11** | 3.8M | 0.02 s |

As can be seen from Table 1, compared with the optimal model GlobalDiff, it seems that GlobalDiff is more competitive in the field of LLIE. The PSNR index of GlobalDiff is better than our model, but in terms of SSIM index, our model performs the same and it is competitive. More importantly, the number of parameters of our model is much less than that of GlobalDiff. The number of parameters is only 21.8% of GlobalDiff and outperforms GlobalDiff in terms of reasoning time. This shows the competitiveness of our model.

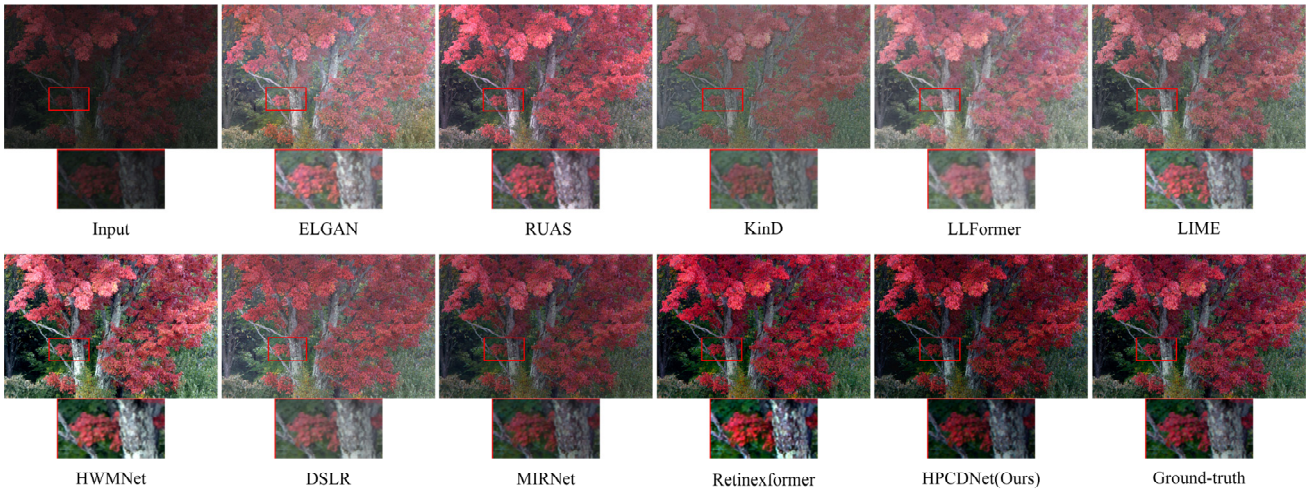


**Figure 7.** In the context of LLIE on the MIT-Adobe FiveK dataset [38], our method demonstrates superior visual results, particularly in the domains of color correction and contrast adjustment.

The intuitive comparison using the MIT-Adobe FiveK dataset is illustrated in Figures 7 and 8. Our method was compared to the current state of the art approach, with results presented in Table 2. Compared to other methods, our approach demonstrates accurate adjustment of image color and contrast, while also showing superior text enhancement. Additionally, our method achieves top-three performance in PSNR and SSIM metrics, surpassed only by MIRNet and Retinexformer, while maintaining outstanding LPIPS results.


**Figure 8.** In comparison to the current state of the art methods on the MIT-Adobe FiveK dataset [38], our approach demonstrates superior performance. It excels in precisely adjusting the image's color and contrast while also outperforming other methods in terms of text enhancement.

**Table 2.** Comparison of low-light enhancement methods performed on the MIT-Adobe FiveK dataset, ↑ (↓) denotes that, larger (smaller) values lead to better performance.

| Method | **PSNR↑** | **SSIM↑** | **LPIP↓** | Method | **PSNR↑** | **SSIM↑** | **LPIP↓** |
|---|---|---|---|---|---|---|---|
| ELGAN [22] | 15.91 | 0.82 | 0.15 | HWMNet [7] | 19.81 | 0.87 | 0.09 |
| RUAS [17] | 16.99 | 0.87 | 0.13 | DSLR [58] | 20.24 | 0.83 | 0.15 |
| KinD [36] | 17.07 | 0.78 | 0.19 | MIRNet [33] | 23.73 | 0.93 | 0.06 |
| LLFormer [30] | 18.75 | 0.84 | 0.15 | Retinexformer [59] | 24.94 | 0.91 | 0.06 |
| LIME [60] | 18.91 | 0.75 | 0.11 | **HPCDNet(Ours)** | 21.97 | 0.90 | 0.05 |

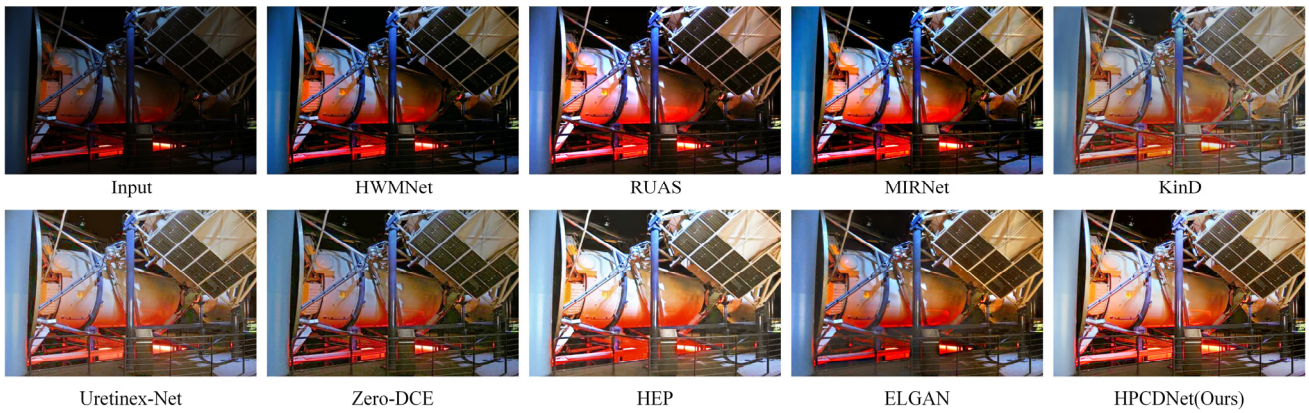


**Figure 9.** Visual comparison on the DICM dataset [47].

We further exhibit the visual enhancement results on the DICM dataset in Figure 9. From the enhancement effect of image illumination, our algorithm is better than other comparative algorithms, which can more naturally brighten the low-light area and show richer details. Meanwhile, the bright area will not show obvious overexposure phenomenon, and the overall color and contrast are balanced to improve and maintain natural tonal consistency with the input image. Other algorithms still have

some deficiencies in the brightening effect, such as the HWMNet, RUAS, and MIRNet algorithms are not enough to brighten the dark areas of the processed image, as well as KinD, Uretinex-Net algorithms that have more noise points, which weaken the visual effect.

### *4.3. Ablation studies*

Ablation studies were performed to assess the impact of four key architectural components on model performance: (1) hybrid position encoding, (2) absolute position encoding, (3) relative position encoding, (4) connection with weights, (5) CFB, and (6) DFAB. The model was trained on 128 × 128 images from the LOL dataset for 600 epochs, with performance evaluated using PSNR. Figure 10 illustrates the visual impacts of ablating each individual component from the model. Quantitative results are tabulated in Table 3. Our ablation study enables several key conclusions:

(1) By introducing hybrid position encoding, the PSNR of the model is improved by 1.05 dB, proving the effectiveness of this module. The improvement of PSNR index without increasing model parameters proves the efficiency of our proposed encoding method. However, due to the extra computation, the training time was extended by 4.01 hours.

(2) Through an independent analysis of absolute position encoding, we performed an ablation study and noted the following key observations: Removing only the absolute position encoding results in a minor 0.69 dB decrease in PSNR, while still improving PSNR by 0.36 dB compared to not using position encoding. Despite no change in model parameters, a slight 0.5 hour increase in training time per 600 epochs was observed relative to the absence of position encoding. This minimal additional computational cost demonstrates the feasibility of incorporating absolute position encoding.

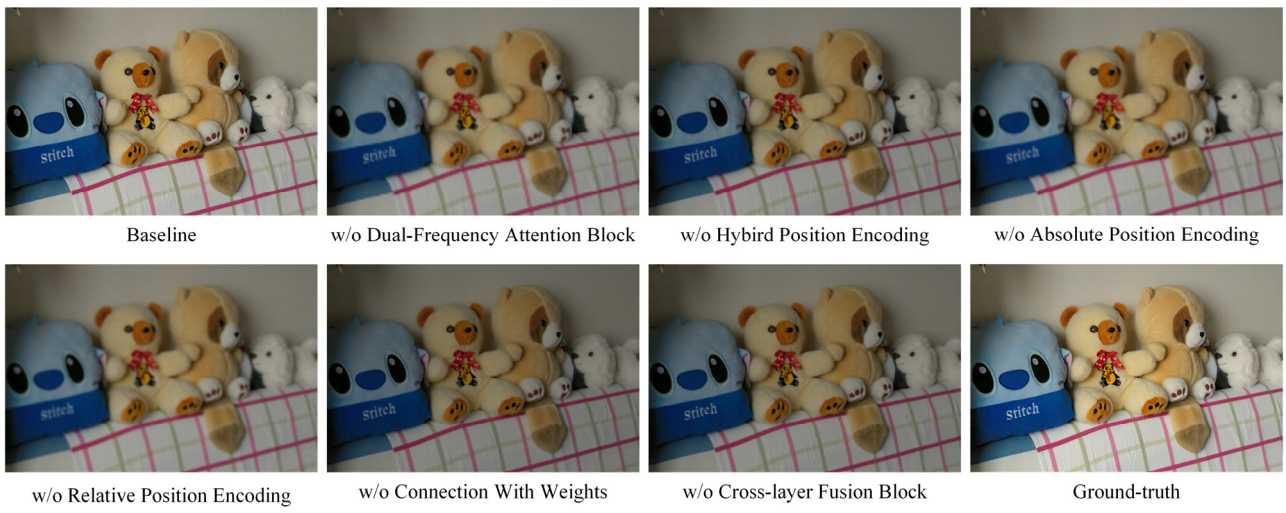


**Figure 10.** Visual comparison of the impact of individual module omissions on our model's performance, highlighting the contributions of key techniques used in our study.

**Table 3.** Ablation experiments were conducted to analyze the contribution of four critical components of the proposed HPCDNet architecture.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Hybrid position encoding | | √ | √ | √ | √ | √ | √ |
| Absolute position encoding | √ | | √ | √ | √ | √ | √ |
| Relative position encoding | √ | √ | | √ | √ | √ | √ |
| Connection with weights | √ | √ | √ | | √ | √ | √ |
| Cross-layer fusion block | √ | √ | √ | √ | | √ | √ |
| Dual-frequency attention block | √ | √ | √ | √ | √ | | √ |
| Params(M) | 3.8 | 3.8 | 3.8 | 3.6 | 3.5 | 3.8 | 3.8 |
| Training time(H) | 13.73 | 14.23 | 15.28 | 13.94 | 14.11 | 16.22 | 17.74 |
| PSNR | 22.18 | 22.64 | 22.31 | 22.94 | 22.94 | 22.36 | 23.23 |

(3) Furthermore, an ablation study was conducted on solely removing the relative position encoding. This resulted in a 0.92 dB decrease in PSNR compared to the full model. Meanwhile, the training time reduced by 2.46 hours relative to the baseline. Though compared to ablating just the absolute position encoding, training time increased by 1.05 hours.

(4) The introduction of connection with weights only achieved a small PSNR improvement of 0.29 dB. At the same time, it increases the parameter size by 0.2 M and introduces additional computational overhead. The performance improvement is small, indicating that the benefit of this module is limited.

(5) Using CFB results in an increase of 0.3 M parameters, and a moderate increase in computational complexity. However, this module plays a key role in representing layered features, improving PSNR by 0.49 dB.

(6) DFAB brings a significant PSNR improvement of 0.87 dB, with minimal computational overhead and no additional parameters, proving that combining the spatial domain and frequency domain is effective.

These discussed module parameters have very little overhead. While individually they may not significantly enhance the network, the additional training time required is negligible. This means we can optimize these factors to improve the model without requiring extensive computing resources. Compared to other algorithms that rely on large models and a large number of parameters, our method requires only few parameters to run. Compared with HWMNet (PSNR: 24.24 dB, Params: 3.6 M), the PSNR of our network is 0.59 dB higher, although the number of parameters is 0.2 M more than HWMNet. Similarly, compared with LLFormer (PSNR: 23.65 dB, Params: 24.5 M), our PSNR is not only 1.18 dB higher than LLFormer, but also reduces the number of parameters by 85%, which greatly reduces the number of parameters of the model. This benefits from the hybrid position encoding, dual frequency domain transformation module and other efficient components we designed, and our model can be embedded into external devices.

## 5. Conclusions

This work proposed an LLIE network. Thanks to the hybrid positional encoding module, the network was able to model both local pixel-level interactions and global dependencies in larger image areas. By applying attention analysis and weighting frequency domain information in the frequency

domain through a dual frequency domain transformation module, the global structure and local details of low-light images were efficiently enhanced. Through the cross-layer fusion module, the features of the previous layer were fused to form high-order features that represent global information, thereby enhancing the expressive ability of the network. Experimental results showed that our network achieves good results on both the LOL dataset and the MIT-Adobe dataset, especially in restoring image details and improving contrast. However, the improvement of this hybrid position encoding for LLIE is limited. In the future, we will continue to explore more efficient position encoding methods to further improve the model's modeling ability of global and local features, thereby achieving greater success. levels and performance improvements. At the same time, we will also study other modules that contribute to image enhancement to enrich the expressive capabilities of the network so that it can generate more natural and clear results.

## Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

## Acknowledgments

This work was supported by the Sichuan Provincial Department of Science and Technology under grant 2022YFS0518 and 2022ZHCG0035, the Open Research Topics of Sichuan International Joint Research Center for Robotics and Intelligent Systems under grant JQZN2022-005, the Opening Fund of Artificial Intelligence Key Laboratory of Sichuan Province under grant 2022RZY05 and Sichuan University of Science and Engineering Postgraduate Innovation Fund in 2023 under grant Y2023312.

## Conflict of interest

The authors declare there is no conflict of interest.

## References

1. M. Chen, Z. Lan, Z. Duan, S. Yi, Q. Su, HDS-YOLOv5: An improved safety harness hook detection algorithm based on YOLOv5s, *Math. Biosci. Eng.*, **20** (2023), 15476–15495. https://doi.org/10.3934/mbe.2023691
2. Y. Wei, Z. Zhang, Y. Wang, M. Xu, Y. Yang, S. Yan, et al., Deraincyclegan: Rain attentive cyclegan for single image deraining and rainmaking, *IEEE Trans. Image Process.*, **30** (2021), 4788–4801. https://doi.org/10.1109/TIP.2021.3074804
3. M. Chen, S. Yi, Z. Lan, Z. Duan, An efficient image deblurring network with a hybrid architecture, *Sensors*, **23** (2023). https://doi.org/10.3390/s23167260
4. M. Abdullah-Al-Wadud, M. Kabir, M. A. Dewan, O. Chae, A dynamic histogram equalization for image contrast enhancement, *IEEE Trans. Consum. Electron.*, **53** (2007), 593–600. https://doi.org/10.1109/TCE.2007.381734
5. D. J. Jobson, Z. Rahman, G. A. Woodell, Properties and performance of a center/surround retinex, *IEEE Trans. Image Process.*, **6** (1997), 451–462. https://doi.org/10.1109/83.557356

6. X. Dong, W. Xu, Z. Miao, L. Ma, C. Zhang, J. Yang, et al., Abandoning the bayer-filter to see in the dark, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 17431–17440. https://doi.org/10.1109/CVPR52688.2022.01691
7. C. M. Fan, T. J. Liu, K. H. Liu, Half wavelet attention on M-Net+ for low-light image enhancement, in *2022 IEEE International Conference on Image Processing (ICIP)*, (2022), 3878–3882. https://doi.org/10.1109/ICIP46576.2022.9897503
8. Z. Cui, K. Li, L. Gu, S. Su, P. Gao, Z. Jiang, et al., You only need 90K parameters to adapt light: A light weight transformer for image enhancement and exposure correction, *BMVC*, **2022** (2022), 238. https://doi.org/10.48550/arXiv.2205.14871
9. S. Moran, P. Marza, S. McDonagh, S. Parisot, G. Slabaugh, Deeplpf: Deep local parametric filters for image enhancement, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 12826–12835. https://doi.org/10.1109/CVPR42600.2020.01284
10. K. Jiang, Z. Wang, Z. Wang, C. Chen, P. Yi, T. Lu, et al., Degrade is upgrade: Learning degradation for low-light image enhancement, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **36** (2022), 1078–1086. https://doi.org/10.1609/aaai.v36i1.19992
11. W. Yang, S. Wang, Y. Fang, Y. Wang, J. Liu, 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), 3063–3072. https://doi.org/10.1109/CVPR42600.2020.00313
12. K. Xu, X. Yang, B. Yin, R. W. Lau, Learning to restore low-light images via decomposition-and-enhancement, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 2281–2290. https://doi.org/10.1109/CVPR42600.2020.00235
13. X. Xu, R. Wang, C. W. Fu, J. Jia, SNR-aware low-light image enhancement, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 17714–17724. https://doi.org/10.1109/CVPR52688.2022.01719
14. C. Wei, W. Wang, W. Yang, J. Liu, Deep retinex decomposition for low-light enhancement, preprint, arXiv:1808.04560. https://doi.org/10.48550/arXiv.2109.05923
15. J. Tan, T. Zhang, L. Zhao, D. Huang, Z. Zhang, Low-light image enhancement with geometrical sparse representation, *Appl. Intell.*, **53** (2022), 1019–1033. https://doi.org/10.1007/s10489-022-04013-1
16. Y. Wang, R. Wan, W. Yang, H. Li, L. P. Chau, A. Kot, Low-light image enhancement with normalizing flow, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2022), 2604–2612. https://doi.org/10.1609/aaai.v36i3.20162
17. R. Liu, L. Ma, J. Zhang, X. Fan, Z. Luo, 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), 10561–10570. https://doi.org/10.1109/CVPR46437.2021.01042
18. W. Yang, W. Wang, H. Huang, S. Wang, J. Liu, Sparse gradient regularized deep retinex network for robust low-light image enhancement, *IEEE Trans. Image Process.*, **30** (2021), 2072–2086. 10.1109/TIP.2021.3050850
19. W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, J. Jiang, 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), 5901–5910. https://doi.org/10.1109/CVPR52688.2022.00581

20. H. Liu, W. Zhang, W. He, Low-light image enhancement based on Retinex theory for beam-splitting prism system, *J. Phys. Conf. Ser.*, **2478** (2023), 062021. https://doi.org/10.1088/1742-6596/2478/6/062021
21. Z. Zhao, B. Xiong, L. Wang, Q. Ou, L. Yu, F. Kuang, RetinexDIP: A unified deep framework for low-light image enhancement, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2021), 1076–1088. https://doi.org/10.1109/TCSVT.2021.3073371
22. Y. F. Jiang, X. Y. Gong, D. Liu, Y. Cheng, C. Fang, X. H. Shen, et al., Enlightengan: Deep light enhancement without paired supervision, *IEEE Trans. Image Process.*, **30** (2021), 2340–2349. https://doi.org/10.1109/TIP.2021.3051462
23. F. Zhang, Y. Shao, Y. Sun, K. Zhu, C. Gao, N. Sang, Unsupervised low-light image enhancement via histogram equalization prior, preprint, arXiv:2112.01766. https://doi.org/10.48550/arXiv.2112.01766
24. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16 x 16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929
25. W. Xu, L. Zou, Z. Fu, L. Wu, Y. Qi, Two-stage 3D object detection guided by position encoding, *Neurocomputing*, **501** (2022), 811–821. 10.1016/j.neucom.2022.06.030
26. M. Tiwari, S. S. Lamba, B. Gupta, A software supported image enhancement approach based on DCT and quantile dependent enhancement with a total control on enhancement level: DCT-Quantile, *Multimedia Tools Appl.*, **78** (2019), 16563–16574. https://doi.org/10.1007/s11042-018-7056-4
27. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, *Adv. Neural Inf. Process. Syst.*, **2017** (2017), 30.
28. Y. Wu, C. Pan, G. Wang, Y. Yang, J. Wei, C. Li, 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*, (2023) 1662–1671.
29. P. Shaw, J. Uszkoreit, A. Vaswani, Self-attention with relative position representations, preprint, arXiv:1803.02155. https://doi.org/10.48550/arXiv.1803.02155
30. T. Wang, K. Zhang, T. Shen, W. Luo, B. Stenger, T. Lu, Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2023), 2654–2662. https://doi.org/10.1609/aaai.v37i3.25364
31. Z. Zhang, Y. Wei, H. Zhang, Y. Yang, S. Yan, M. Wang, Data-driven single image deraining: A comprehensive review and new perspectives, *Pattern Recognit.*, **2023** (2023), 109740. https://doi.org/10.1016/j.patcog.2023.109740
32. S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M. H. Yang, Restormer: Efficient transformer for high-resolution image restoration, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 5728–5739. https://doi.org/10.1109/CVPR52688.2022.00564
33. S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M. H. Yang, et al., Learning enriched features for fast image restoration and enhancement, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2023), 1934–948. https://doi.org/10.1109/TPAMI.2022.3167175

34. K. G. Lore, A. Akintayo, S. Sarkar, LLNet: A deep autoencoder approach to natural low-light image enhancement, *Pattern Recognit.*, **61** (2017), 650–662. https://doi.org/10.1016/j.patcog.2016.06.008
35. Y. Zhang, X. Guo, J. Ma, W. Liu, J. Zhang, Beyond brightening low-light images, *Int. J. Comput. Vision*, **129** (2021), 1013–1037. https://doi.org/10.1007/s11263-020-01407-x
36. Y. Zhang, J. Zhang, X. Guo, Kindling the darkness: A practical low-light image enhancer, in *Proceedings of the 27th ACM International Conference on Multimedia*, (2019), 1632–1640. https://doi.org/10.1145/3343031.3350926
37. Z. Zhang, H. Zheng, R. Hong, M. Xu, S. Yan, M. Wang, Deep color consistent network for low-light image enhancement, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 1899–1908. https://doi.org/10.1109/CVPR52688.2022.00194
38. Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, R. Salakhutdinov, Transformer-xl: Attentive language models beyond a fixed-length context, preprint, arXiv:1901.02860. https://doi.org/10.48550/arXiv.1901.02860
39. Z. Huang, D. Liang, P. Xu, B. Xiang, Improve transformer models with better relative position embeddings, preprint, arXiv:2009.13658. https://doi.org/10.48550/arXiv.2009.13658
40. P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, Stand-alone self-attention in vision models, *Adv. Neural Inf. Process. Syst.*, **2019** (2019), 32.
41. H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, L. C. Chen, Axial-deeplab: Stand-alone axial-attention for panoptic segmentation, in *European Conference on Computer Vision*, (2020), 108–126. https://doi.org/10.1007/978-3-030-58548-8_7
42. K. Wu, H. Peng, M. Chen, J. Fu, H. Chao, Rethinking and improving relative position encoding for vision transformer, in P*roceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 10033–10041.
43. N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, et al., Image transformer, in *International Conference on Machine Learning: PMLR*, (2018), 4055–4064.
44. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in *European Conference on Computer Vision*, (2020), 213–229. https://doi.org/10.1007/978-3-030-58452-8_13
45. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, *Adv. Neural Inf. Process. Syst.*, **34** (2021), 12077–12090.
46. D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), preprint, arXiv:1606.084150. https://doi.org/10.48550/arXiv.1606.08415
47. S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M. H. Yang, et al., Cycleisp: Real image restoration via improved data synthesis, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 2696–2705. https://doi.org/10.1109/CVPR42600.2020.00277
48. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 7132–7141.
49. F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, et al., Residual attention network for image classification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 3156–3164. https://doi.org/10.1109/CVPR.2017.683

50. M. Jaderberg, K. Simonyan, A. Zisserman, Spatial transformer networks, *Adv. Neural Inf. Process. Syst.*, **2015** (2015), 28.
51. I. Daubechies, Orthonormal bases of compactly supported wavelets, *Commun. Pure Appl. Math.*, **41** (1988), 909–996. https://doi.org/10.1002/cpa.3160410705
52. K. R. Rao, P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*, Academic press, 2014. https://doi.org/10.1016/c2009-0-22279-3
53. Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, H. Li, Uformer: A general u-shaped transformer for image restoration, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 17683–17693. https://doi.org/10.1109/CVPR52688.2022.01716
54. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556
55. T. Wang, K. Zhang, Z. Shao, W. Luo, B. Stenger, T. K. Kim, et al., LLDiffusion: Learning degradation representations in diffusion models for low-light image enhancement, preprint, arXiv:2307.14659. https://doi.org/10.48550/arXiv.2307.14659
56. J. Hou, Z. Zhu, J. Hou, H. Liu, H. Zeng, H. Yuan, Global structure-aware diffusion process for low-light image enhancement, preprint, arXiv:2310.17577. https://doi.org/10.48550/arXiv.2310.17577
57. X. Yi, H. Xu, H. Zhang, L. Tang, J. Ma, Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2023), 12302–12311.
58. S. Lim, W. Kim, DSLR: Deep stacked Laplacian restorer for low-light image enhancement, *IEEE Trans. Multimedia,* **23** (2020), 4272–4284. https://doi.org/10.1109/TMM.2020.3039361
59. Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, Y. Zhang, Retinexformer: One-stage Retinex-based transformer for low-light image enhancement, preprint, arXiv:2303.06705. https://doi.org/10.48550/arXiv.2303.06705
60. X. Guo, Y. Li, H. Ling, LIME: Low-light image enhancement via illumination map estimation, *IEEE Trans. Image Process.*, **26** (2016), 982–993. https://doi.org/10.1109/TIP.2016.2639450



AIMS Press