

# Holistic Dynamic Frequency Transformer for Image Fusion and Exposure Correction

Xiaoke Shang<sup>a</sup>, Gehui Li<sup>b</sup>, Zhiying Jiang<sup>c,d</sup>, Shaomin Zhang<sup>a</sup>, Nai Ding<sup>a</sup> and Jinyuan Liu<sup>e,\*</sup>

<sup>a</sup>School of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou 310027, China

<sup>b</sup>DUT-RU International School of Information Science & Engineering, Dalian University of Technology, Dalian 116620, China

<sup>c</sup>School of Software Technology, Dalian University of Technology, Dalian 116024, China

<sup>d</sup>Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian 116620, China

<sup>e</sup>School of Mechanical Engineering, Dalian University of Technology, Dalian 116086, China

## ARTICLE INFO

### Keywords:

Exposure correction  
Low-light enhancement  
Multi-exposure fusion  
Low-light face detection  
Low-light segmentation

## ABSTRACT

The correction of exposure-related issues is a pivotal component in enhancing the quality of images, offering substantial implications for various computer vision tasks. Historically, most methodologies have predominantly utilized spatial domain recovery, offering limited consideration to the potentialities of the frequency domain. Additionally, there has been a lack of a unified perspective towards low-light enhancement, exposure correction, and multi-exposure fusion, complicating and impeding the optimization of image processing. In response to these challenges, this paper proposes a novel methodology that leverages the frequency domain to improve and unify the handling of exposure correction tasks. Our method introduces Holistic Frequency Attention and Dynamic Frequency Feed-Forward Network, which replace conventional correlation computation in the spatial-domain. They form a foundational building block that facilitates a U-shaped Holistic Dynamic Frequency Transformer as a filter to extract global information and dynamically select important frequency bands for image restoration. Complementing this, we employ a Laplacian pyramid to decompose images into distinct frequency bands, followed by multiple restorers, each tuned to recover specific frequency-band information. The pyramid fusion allows a more detailed and nuanced image restoration process. Ultimately, our structure unifies the three tasks of low-light enhancement, exposure correction, and multi-exposure fusion, enabling comprehensive treatment of all classical exposure errors. Benchmarking on mainstream datasets for these tasks, our proposed method achieves state-of-the-art results, paving the way for more sophisticated and unified solutions in exposure correction.

## 1. Introduction

Exposure-related tasks, including low-light enhancement, exposure correction, and multi-exposure fusion, bear significant implications in the field of computer vision. Proper image exposure is instrumental in achieving high-quality visual information, making it pivotal for effective image analysis and processing. Overexposure and underexposure, both common imaging issues, lead to loss of details and reduced contrast, thereby hindering the visual appeal and practical usability of images.

The evolution of methodologies to address low-light enhancement has seen significant advancements over the years. Initial methods relied on traditional image processing techniques, including histogram equalization and gamma correction. These were soon replaced by sophisticated techniques leveraging deep learning. Notable among these are Deep-UPE [1] and Zero-DCE [2], which optimized for local contrast enhancement in underexposed images. However, these models primarily focused on underexposure correction, leaving overexposed images relatively unaddressed. This led to the expansion of low-light enhancement techniques into exposure correction tasks, an evolution marked by the introduction of the Multi-Scale Exposure Correction

(MSEC) [3] method, which targets both over- and underexposed regions in images. LCDP [4] is proposed to rectify multiple exposure errors present within the same image. FECNet [5], a Fourier and convolution-based Exposure Correction Network, comprises an amplitude sub-network and a phase sub-network, which restore the amplitude and phase representations respectively. Existing exposure correction networks have not adequately addressed the issues of color and detail loss during the correction process. Moreover, a comprehensive integration and application of Transformer networks with frequency domain approaches have not been explored. The task generalization of existing schemes also remains a pressing issue that needs to be resolved.

As for the multi-exposure fusion tasks, the prevailing solutions can be generally divided into several categories. Traditional methods [6, 7] like gradient-based fusion and pyramid-based fusion provide a foundation but often result in artifacts or insufficient detail preservation. More recent contributions have capitalized on the power of deep learning, such as the fusion algorithm based on a convolutional neural network [8–14]. While they advanced the performance of fusion tasks significantly, they still lacked a holistic approach to addressing exposure-related tasks in their entirety.

Our proposed method intervenes to overcome existing limitations by introducing an innovative, unified framework for exposure correction that spans the domains of low-light enhancement, exposure correction, and multi-exposure fusion. Recognizing that low-light data serves as a subset of

\*Corresponding author:

✉ atlantis918@hotmail.com (Jinyuan Liu)

ORCID(s): 0000-0003-2085-2676 (Jinyuan Liu)

exposure-error data, which includes both overexposure and underexposure, and multi-exposure fusion often necessitates subsequent detail and color restoration and enhancement, our approach centers on devising a robust, universal exposure corrector engineered to harmonize these three intertwined tasks into a cohesive solution.

We propose Holistic Frequency Attention (HFA), and Dynamic Frequency Feed-Forward Network (DFFFN). They form a U-shaped restorer that establishes global dependencies and dynamically filters the main information according to the frequency band. Paired with our Laplacian pyramid for image decomposition and pyramid fusion, the approach offers fine-grained and nuanced image restoration, resulting in enhanced detail and color correction.

The contributions of this paper are as follows:

- We introduce a U-shaped Restorer that combines Holistic Frequency Attention and Dynamic Frequency Feed-Forward Network, designed to efficiently extract global features and dynamically filter crucial frequency bands, thereby achieving enhanced detail and color in exposure-corrected images.
- Our method leverages Laplacian pyramids to decompose images and employs multiple restorers for pyramid fusion, enhancing the information synthesis from different frequency bands.
- We unveil an integrated architecture that synergistically addresses low-light enhancement, exposure correction, and multi-exposure fusion tasks by leveraging the inherent correlations among these three challenges, demonstrating its versatility in tackling a wide range of exposure-related issues.
- Our method achieves state-of-the-art performance on both downstream and upstream tasks, demonstrating its effectiveness and applicability.

## 2. Related Work

### 2.1. Low-Light Enhancement and Exposure-Error Correction

Low-light image enhancement has been an active research topic in computer vision due to its importance in improving the visual quality of images and the performance of visual recognition tasks under poor lighting conditions. Methods for low-light image enhancement can be broadly categorized into three types: histogram-based methods, Retinex-based methods, and deep learning-based methods.

Initially, research predominantly focused on histogram-based methods, which offered a computationally efficient and straightforward approach. These methods typically manipulate the histogram of input images, stretching the contrast to boost the visibility of features under low-light conditions. Seminal works like the Dynamic Histogram Equalization (DHE) by Abdullah and Fofi [15], and the Adaptive Histogram Equalization (AHE) by Pizer et al. [16], exemplify this approach. In parallel, Reza [17] proposed an

efficient realization of histogram equalization, while Ying et al. [18] introduced the contrast-limited adaptive histogram equalization (CLAHE) technique, specifically designed to suppress noise over-amplification.

Building upon these foundational methods, subsequent researchers began exploring the potential of the Retinex theory. This theory involves the separation of an image into illumination and reflectance components, enabling finer control over image enhancement. Influential contributions include the Retinex model by Land and McCann [19], and the low-light image enhancement (LIME) method by Guo et al. [20]. Over time, more sophisticated Retinex-based models were developed, such as the robust Retinex model by Li et al. [21], the low-rank regularized Retinex model (LR3M) by Ren et al. [22], and the joint intrinsic-extrinsic prior model by Cai et al. [23].

With the advent of deep learning, researchers have found a potent tool for tackling low-light enhancement [24, 25]. Chen et al. [26] effectively harnessed two-way generative adversarial networks (GANs) in an unpaired learning method for image enhancement. In a similar vein, Lv et al. [27] utilized a multi-branch convolutional neural network in an end-to-end attention-guided approach. Other research, such as that by Wang et al. [1], explored neural networks' capabilities in enhancing underexposed photos by introducing intermediate illumination into the network. Wei et al. [28] proposed unique solutions by integrating signal prior-guided layer separation, data-driven mapping networks, and deep Retinex-Nets for low-light enhancement.

Recent studies in the field have expanded beyond low-light enhancement to address exposure correction [29], tackling both over- and under-exposed images. Afifi et al. [3] laid the groundwork by splitting the exposure correction into color and detail enhancement tasks and using a wide-ranging exposure dataset. Building on this, Cui et al. [30] developed the Illumination Adaptive Transformer (IAT) to adjust color correction and gamma correction parameters in images under various lighting conditions. Complementing these efforts, Wang et al. [4] proposed a method exploiting local color distributions to enhance regions suffering from both over- and under-exposure, introducing a dual-illumination learning mechanism and a new dataset to aid this process.

### 2.2. Multi-Exposure Fusion

The development of Multi-Exposure Image Fusion [31–45] has seen many significant strides, primarily utilizing deep learning methodologies. Kalantari et al. [46] utilized a Convolutional Neural Network (CNN) to merge HDR images, mitigating ghosting and tearing artifacts often appearing in dynamic scenes. Contrastingly, Ma et al. [47] proposed MEF-Net, a swift MEF method employing a fully convolutional network for weight map prediction, outpacing and outperforming many traditional methods. Yin et al. [48] integrated both pixel-level and feature-level considerations into a novel encoder-decoder network, ensuring fine-grained control, semantic consistency, and texture calibration. Ram



**Fig. 1:** The proposed method produce a single-exposure-corrected(SEC)/multiple-exposure-fused(MEF) image with clear details and visually pleasing colors.

et al. [49] tackled the limitations of hand-crafted feature-based MEF methods by introducing an unsupervised deep learning architecture, demonstrating superior performance on natural images. Xu et al. [50] introduced MEF-GAN, an adversarial network incorporating self-attention mechanism, allowing for effective correction of local distortion and inappropriate representation. Chen et al. [51] proposed a network that integrates multiple mechanisms such as homography estimation, attention mechanism, and adversarial learning to address ghosting issues and misalignment. Most recently, Liu et al. [52] developed an attention-guided global-local adversarial learning network, ensuring the alignment of local patches of the fused images with realistic normal-exposure ones, thereby restoring realistic texture details and correcting color distortion.

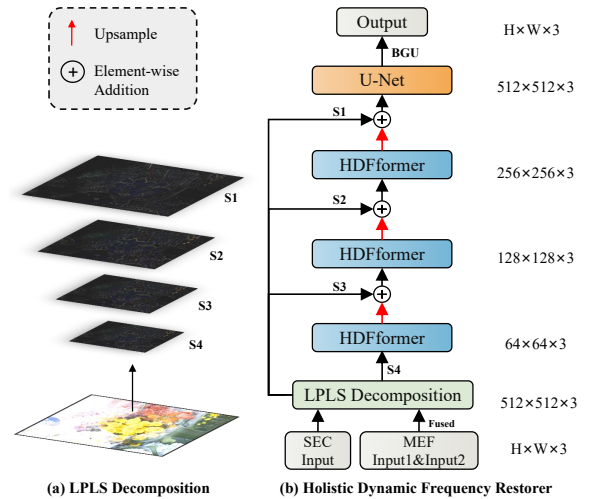
Existing methodologies in the field of Multi-Exposure Image Fusion (MEF) often separately study exposure correction and multi-exposure fusion, with no single method unifying these three tasks. Recognizing the correlation between multi-exposure fusion and exposure correction, our proposed method brings these two aspects under a unified network. By concurrently addressing these tasks, our approach seeks to leverage the interdependencies between them, ultimately aiming to improve the overall performance and efficacy of multi-exposure image fusion.

### 2.3. Fourier Transform in Computer Vision

Fourier Transform has played a pivotal role in advancing computer vision tasks [53–57]. By transferring an image from its spatial domain to the frequency domain, it allows for the identification and extraction of informative features that are often more resilient to local variations and noise. Rao et al. [58] introduced the Global Filter Network (GFNet), a model that learns spatial dependencies in the frequency domain using 2D discrete Fourier transforms, demonstrating excellent accuracy and efficiency. Xu et al. [59] presented a Fourier-based approach for domain generalization tasks. By leveraging Fourier phase information, they developed an effective data augmentation strategy and a co-teacher regularization technique. Chi et al. [60] proposed the Fast Fourier Convolution (FFC) operator, an innovative design

encapsulating different scales of computations within a single unit. Kong et al. [61] proposed frequency domain attention utilizing Fast Fourier Transform to reduce attention complexity, and employed a gated discriminative filtering mechanism to address the deblurring problem.

Despite the success of Fourier Transform in numerous computer vision tasks, there is a noticeable gap in its application for exposure correction and multi-exposure fusion tasks. These tasks have yet to be deeply researched from the frequency domain perspective.



**Fig. 2:** The workflow of the proposed Restorer.

## 3. Proposed Method

We utilize Laplacian pyramids to decompose the images into different frequency bands, then employ multiple restorers, each of which restores information specific to a particular frequency band, thereby facilitating pyramid fusion. This multi-level decomposition and fusion strategy enables more detailed and comprehensive extraction and synthesis of image information from different frequency bands, enhancing the quality of the output images. Figure 2 shows the workflow of the proposed method. In the following, we present the details of each component.

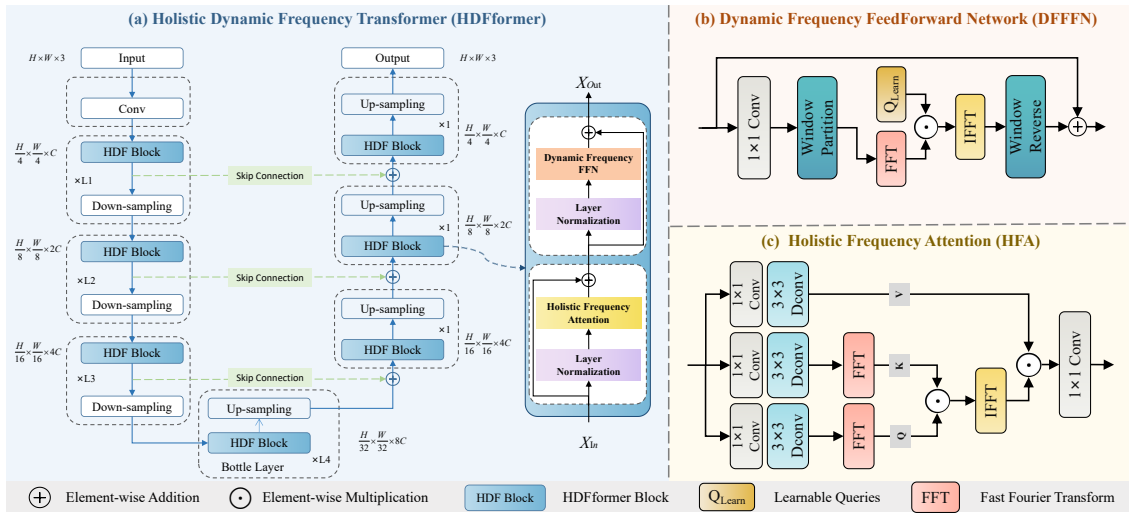


Fig. 3: The workflow of the proposed Holistic Dynamic Frequency Transformer.

### 3.1. Laplacian Pyramid Decomposition

The principle of image pyramid is that each image is decomposed into  $N$  layers of multi-scale pyramid image sequence. The image with the smallest size is taken as the 1st layer and the image with the largest size is taken as the  $N$ th layer. The size of the image in the  $K$ th layer is one-fourth of the size of the image in the  $K+1$ th layer. During the operation of Gaussian pyramid, the image is filtered by Gaussian blur and down-sampling operation will lose some high frequency detail information. To describe this high-frequency information, Laplace pyramid is defined. The image of each layer of the Gaussian pyramid is subtracted from the image of the next layer after up-sampling and Gaussian filtering. A series of difference images are obtained, which are the images after the Laplacian pyramid decomposition [3]. Mathematically defined as:  $L_i = G_i - \text{PyrUp}(G_{i+1})$ . The purpose of this operation is to decompose the source images into different spatial frequency bands, so that separate networks can be used to restore features and details of specific frequency bands at different decomposition layers.

### 3.2. Holistic Dynamic Frequency Transformer

By leveraging multiplication in the frequency domain to replace correlation calculation in the temporal domain, we introduce a novel attention mechanism and FFN layer that are based on the frequency domain. These two components form a building block, which then constitutes a U-shaped restorer, functioning as a filter to selectively restore the required frequency domain for image reconstruction. This innovative concept enhances the computational efficiency and performance of the attention mechanism in computer vision tasks. Figure 3 shows the workflow of the proposed restorer.

#### 3.2.1. Holistic Frequency Attention

Our work primarily focuses on the Window Attention mechanism in the context of computer vision. Window Attention is a variant of the attention mechanism where computations are restricted within local windows, which are subparts of the input image or feature map.

In the Attention mechanism, input  $X$  is multiplied by three mapping matrices  $W_q$ ,  $W_k$ , and  $W_v$  to generate the query  $\mathbf{Q}$ , key  $\mathbf{K}$ , and value  $\mathbf{V}$  respectively:

$$\mathbf{Q} = XW_q, \quad \mathbf{K} = XW_k, \quad \mathbf{V} = XW_v. \quad (1)$$

The formulation of Attention is as follows:

$$A = \text{Softmax}(\mathbf{Q}\mathbf{K}^T / \sqrt{d_k})\mathbf{V}, \quad (2)$$

where  $d_k$  is the dimensionality of the query and key vectors, and  $A$  is the output attention map.

The time complexity of this operation is  $O(N^2)$ , where  $N$  is the number of pixels in the image. The limitation of attention is its computational complexity. When the image size increases, the required computations increase quadratically. Window attention is used to reduce complexity. Its complexity is  $O(DN)$ , where  $D$  is the number of pixels in the window. On the one hand, its complexity deteriorates to quadratic as the window increases. On the other hand, the partitioning into windows can result in loss of long-range dependencies because the attention is only applied within individual windows.

An inspired insight from [61] leverages the classical Convolution Theorem, demonstrating that multiplication in the frequency domain can effectively replace correlation operations in the spatial domain. Accordingly, we propose to replace the matrix multiplication between attention query and key with multiplication in the frequency domain without windowing. Furthermore, from the perspective of [60], this can be interpreted as query and key mutually filtering each other in the frequency domain. The specific process is as follows: First, we use the Fast Fourier Transform (FFT) to transform the query and key into the frequency domain, represented as  $Q_f$  and  $K_f$ :

$$Q_f = FFT(\mathbf{Q}), \quad K_f = FFT(\mathbf{K}). \quad (3)$$

We then perform element-wise multiplication in the frequency domain:

$$M_f = Q_f \odot K_f, \quad (4)$$

where  $\odot$  denotes element-wise multiplication. Subsequently, we use the Inverse Fast Fourier Transform (IFFT) to transform back to the spatial domain:

$$M = IFFT(M_f). \quad (5)$$

The attention matrix  $A$  and value  $V$  are then subject to a Hadamard product, giving the final output after attention:

$$A = Softmax(M) \odot V. \quad (6)$$

Finally, the attention map  $A$  is convolved by a  $1 \times 1$  convolution and added to the original input  $X$ :

$$X' = Conv_{1 \times 1}(A) + X. \quad (7)$$

The overall complexity of our method is  $O(N \log N)$ , which significantly reduces the computational cost compared to traditional Window Attention.

Our method maintains the advantages of local window-based attention, such as preserving local structures and reducing memory requirements, while overcoming its limitations by effectively capturing long-range dependencies and lowering computational complexity.

### 3.2.2. Dynamic Frequency FeedForward Network

To address the issue that not all low and high frequency information are beneficial for effective image recovery, we propose an adaptive method through a Frequency Filtering Network (FFN). The key challenge is how to effectively determine which frequency information is crucial. Inspired by the JPEG compression algorithm, we introduce a learnable querying mechanism  $Q_{Learned}$ . Similar to [58, 60, 61], this approach employs a gating mechanism between the FFT and IFFT operations for filtering. The method differs in its simplicity, directly using a single self-learned parameter to apply identical filtering across all windows. The essence of the learnable query lies in a customizable matrix. It essentially acts as a self-learned prompt, influenced by backpropagation, to learn appropriate parameters from the dataset itself for filtering the frequency domain representations of feature maps. The steps of the proposed method are represented as follows:

---

#### Algorithm 1 Frequency Filtering Network Process

---

- 1:  $I' \leftarrow \Phi(I)$   $\triangleright$   $1 \times 1$  Convolution to increase channels
  - 2:  $W_{original} \leftarrow \omega(I')$   $\triangleright$  Window partition
  - 3:  $W_{freq} \leftarrow \mathcal{F}(W_{original})$   $\triangleright$  FFT to frequency domain
  - 4:  $W_{filtered} \leftarrow W_{freq} \odot Q_{Learned}$   $\triangleright$  Hadamard product
  - 5:  $W_{spatial} \leftarrow \mathcal{F}^{-1}(W_{filtered})$   $\triangleright$  IFFT to spatial domain
  - 6:  $I'' \leftarrow \rho(W_{spatial})$   $\triangleright$  Restoration of image from windows
  - 7:  $O \leftarrow \Phi^{-1}(I'')$   $\triangleright$   $1 \times 1$  Convolution to decrease channels
- 

Here,  $\Phi$  represents the  $1 \times 1$  convolution operation for expanding the number of channels,  $\omega$  stands for the operation of dividing the image into windows,  $\mathcal{F}$  denotes the Fast Fourier Transform (FFT),  $\mathcal{F}^{-1}$  is the Inverse Fast Fourier Transform (IFFT),  $\rho$  stands for the restoration of windows

back into a single image, and  $\Phi^{-1}$  signifies the  $1 \times 1$  convolution operation for reducing the number of channels. By employing this sequence of operations, our method ensures that only the necessary frequency information is preserved, leading to a more effective image recovery.

### 3.3. Image Fusion Block

Our approach adopts a convolutional pathway, transforming the input source image into a distinctive feature representation using convolutional layers. This is mathematically represented as follows:

$$F = Conv(\mathbf{I}), \quad (8)$$

where  $F$  is the resultant feature representation,  $Conv$  represents the multiple convolution operations, and  $\mathbf{I}$  is the input source image. Subsequently, an amalgamation of Max Pooling and Average Pooling is employed for downsampling. The outcome of this is concatenated to realize a perception at various scales.

$$F' = Concat(AvgPool(F), MaxPool(F)), \quad (9)$$

where  $F'$  is the downscaled feature representation. During the processing of low-scale feature maps, we address feature displacement using skip connections during the upsampling phase. The corrected features are then convolved to generate an attention map,

$$A = Conv(Connect(UpSample(F'), F)). \quad (10)$$

In the final stage, the attention map is used to generate two images, which are element-wise multiplied with the corresponding image. The summation of these results forms the initial fused image,

$$Fused_{initial} = \sum_{i=1}^2 A_i \odot I_i. \quad (11)$$

In this scheme, we are able to generate an image, imbued with complementary information extracted from the source image. Nevertheless, it's critical to note that while the attention mechanism retains and integrates a wealth of information from different exposure levels, it does not account for color calibration and exposure correction. Therefore, it is imperative to execute subsequent processing on the initial fused image.

### 3.4. Loss Function

In our proposed method, two types of losses have been incorporated, each contributing to the overall loss function proportionally with respect to their weights. The overall loss function is as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{mse} + \lambda_2 \mathcal{L}_{pyr}. \quad (12)$$

**Reconstruction Loss.** The first component of the overall loss function is the reconstruction loss. The aim of the reconstruction loss is to measure the dissimilarity between

the ground truth image and the image reconstructed by our method. We employ the Mean Squared Error (MSE) as our metric for this. The formula for the reconstruction loss is defined as follows:

$$\mathcal{L}_{mse} = \sum_{p=1}^{3hw} |\mathbf{O}^1(p) - \mathbf{G}(p)|, \quad (13)$$

where  $\mathbf{O}^1$  is the final output image corrected by our method, and  $\mathbf{G}$  is the ground truth image. Here,  $p$  denotes the pixel index, with  $h$  and  $w$  being the height and width of the output image, respectively.

**Laplacian Pyramid Loss.** The second component of our overall loss function is the Laplacian pyramid loss. This is used to measure the dissimilarity between the ground truth image and the image produced by our method on different levels of the Gaussian pyramid. The formula for the Laplacian pyramid loss is:

$$\mathcal{L}_{pyr} = \sum_{i=2}^n 2^{(i-2)} \sum_{p=1}^{3h_i w_i} |\mathbf{O}^i(p) - \mathbf{G}^i(p)|, \quad (14)$$

where  $\mathbf{O}^i$  is the output of the restoration stage at the  $i$ th level of the pyramid, and  $\mathbf{G}^i$  is the  $i$ th level of the Gaussian pyramid, formed from the ground truth image. Here,  $p$  denotes the pixel index,  $h_i$  and  $w_i$  are the height and width of the image at the  $i$ th level of the pyramid, respectively, and  $n$  is the total number of levels in the pyramid.

In the overall loss function,  $\lambda_1$  and  $\lambda_2$  are the weights associated with the reconstruction loss and Laplacian pyramid loss, respectively.

## 4. Experiment

In our research, we endeavor to benchmark our approach against the state-of-the-art methods for exposure correction, low-light enhancement, and multi-exposure fusion, utilizing several classic datasets for both quantitative and qualitative comparison. Our method's applicability is further demonstrated by extending it to high-level tasks. In a bid to corroborate the versatility of our proposed method, we utilized our exposure correction technique as a pre-enhancer for low-light images. This preparatory step significantly improved the performance of subsequent high-level tasks such as low-light face detection and low-light semantic segmentation. The performance metrics of these tasks serve to underscore the robustness of our approach in real-world applications, bridging the gap between low-level image enhancement and high-level visual understanding.

### 4.1. Datasets for Exposure-related Tasks

In our study, we thoroughly evaluated our proposed method across five distinct tasks: exposure correction, low-light enhancement, multi-exposure fusion, low-light face detection, and low-light semantic segmentation. In the following sections, we detail the datasets selected for this evaluation, outlining their key characteristics and the reasons for their inclusion in our experimental setup.

#### 4.1.1. Datasets for Exposure Correction

Our experiments involve two exposure-errors datasets, MSEC and LCDP. MSEC dataset contains 24,330 8-bit sRGB images divided into 17,675 training images, 750 validation images, and 5905 test images. The images in them are tuned by the MIT-Adobe FiveK dataset with 5 different exposure values (EV) ranging from under-exposure to over-exposure conditions. Each image of the training set is accompanied by a ground truth image. Each image of the test set has manual correction results from 5 different experts (A/B/C/D/E). The LCDP has 1733 pairs of images, which are divided into 1415 pairs for training, 100 pairs for validation, and 218 pairs for testing. Each image in the MSEC dataset has an overall exposure error, while each image in LCDP has different types of exposure errors in different regions. These two datasets represent the types of exposure errors commonly found in reality and are the only two large-scale datasets available for the exposure correction.

#### 4.1.2. Datasets for Low-light Enhancement

The LOL dataset, introduced by Wei et al., is specifically designed for low-light image enhancement. It comprises 500 image pairs, each consisting of a low-light image and its corresponding normally lit image. The images in the dataset are categorized into two types: 400 image pairs are selected from the web and serve as the training set, while the remaining 100 pairs, captured by various smartphones in real-life low-light scenarios, form the test set. The LOL dataset provides a challenging and practical environment for low-light enhancement algorithms, as it covers a diverse range of scenarios, such as indoor, outdoor, dawn, night and backlit scenes, effectively mimicking real-world conditions.

The MIT-Adobe FiveK dataset, introduced by Bychkovsky et al., is a large-scale dataset originally designed for color enhancement and editing. It consists of 5000 high-quality RAW photographs, each retouched by five different photographers, providing a total of 25000 enhanced images. For the purpose of low-light enhancement, a subset of this dataset is commonly used. This subset includes images captured under various challenging lighting conditions, such as at sunset, under cloudy weather, or indoors with artificial lighting. Each image in this subset is accompanied by multiple retouched versions, offering multiple possible 'ground truths' and thus promoting a more comprehensive evaluation of the enhancement methods.

#### 4.1.3. Datasets for Multi-exposure Fusion

In our research, we cultivated a specific subset of 490 image sequences from the SICE dataset, each sequence consisting of an over-exposed, under-exposed, and a high-quality reference image, handpicked to represent extreme exposure scenarios. A set of 360 sequences was randomly chosen for training, with the remaining 130 sequences utilized for validation. For experimental comparison, we adopted 100 randomly picked image pairs from the SICE dataset, supplemented with an additional 18 pairs without ground truth,

thus gauging the versatility of our method under various conditions.

#### 4.1.4. Baseline and Datasets for High-Level Tasks

We utilized the S3FD[62], a well-known face detection algorithm to evaluate the dark face detection performance and adopted the PSPNet[63] as the baseline to evaluate the segmentation performance. Low-light human face detection and low-light semantic segmentation are experimented on the DARKFACE [64] and ACDC [65] datasets respectively. To enhance adaptability, we fine-tuned all visual models on the corrected output.

The DARKFACE dataset [64] is designed for low-light human face detection. It includes 6,000 low-light images from various real-world settings, labeled with bounding boxes identifying human faces. Additionally, the dataset consists of 9,000 unlabeled low-light images and a unique subset of 789 images captured under both low-light and normal lighting conditions. A hold-out testing set of 4,000 low-light images, annotated with human face bounding boxes, is also provided.

The ACDC dataset [65] is used for semantic segmentation under adverse conditions. It comprises 4,006 images, equally distributed under four common adverse conditions: fog, nighttime, rain, and snow. Each image in the dataset is accompanied by fine pixel-level semantic annotations and a binary mask. This mask differentiates between regions of clear and uncertain semantic content within the image, supporting both standard semantic segmentation and uncertainty-aware semantic segmentation.

## 4.2. Metrics for Image Quality Assessment

Image quality assessment is a fundamental aspect of various processes in computer vision and image processing. The reliability and effectiveness of the proposed methodologies are quantitatively evaluated based on several key metrics. This paper particularly leverages the Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Multi-Exposure Fusion Structural Similarity Index (MEF-SSIM) to gauge the quality of processed images and thereby ascertain the performance of the proposed approach.

### 4.2.1. Peak Signal to Noise Ratio (PSNR)

The Peak Signal to Noise Ratio (PSNR) is a commonly used objective metric for assessing the quality of reconstruction of lossy compression codecs for image and video data. It is a simple yet effective measure of the error between a reference image and a distorted version of the image.

The PSNR is formally defined as:

$$\text{PSNR} = 10 \cdot \log_{10} \left( \frac{\text{MAX}_I^2}{\text{MSE}} \right), \quad (15)$$

where  $\text{MAX}_I$  is the maximum possible pixel value of the image. For an 8-bit grayscale image, the maximum pixel value is 255. MSE is the Mean Squared Error, which is

the average squared difference between the pixels of the reference image and the distorted image.

### 4.2.2. Structural Similarity Index Measure (SSIM)

The Structural Similarity Index Measure (SSIM) is a perception-based model that considers changes in structural information, illumination, and contrast as separate components that contribute to the quality of an image.

The SSIM index is calculated as:

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \quad (16)$$

where  $\mu_x$  and  $\mu_y$  are the average of  $\mathbf{x}$  and  $\mathbf{y}$ ;  $\sigma_x^2$  and  $\sigma_y^2$  are the variance of  $x$  and  $y$ ;  $\sigma_{xy}$  is the covariance of  $x$  and  $y$ ;  $c_1 = (k_1 L)^2$ ,  $c_2 = (k_2 L)^2$  are two variables to stabilize the division with weak denominator;  $L$  is the dynamic range of pixel-values;  $k_1 = 0.01$  and  $k_2 = 0.03$  by default.

### 4.2.3. Multi-Exposure Fusion Structural Similarity Index Measure (MEF-SSIM)

The Multi-Exposure Fusion Structural Similarity Index is an extension of the traditional SSIM that is specifically designed for evaluating the quality of images fused from multiple exposures. It takes into account the quality of the details in the fused image, the naturalness of the fused image, and the visibility of the image content in different exposures.

The MEF-SSIM index is calculated as:

$$\text{MEF-SSIM}(\mathbf{I}_f, \mathbf{I}_i) = \text{SSIM}(\mathbf{I}_f, \mathbf{I}_i) \cdot W_i, \quad (17)$$

where  $\mathbf{I}_f$  is the fused image and  $\mathbf{I}_i$  is the  $i$ -th source image.  $W_i$  is the weight of the  $i$ -th source image, calculated as:

$$W_i = \frac{\exp(-\beta \cdot (\mathbf{I}_i - \bar{I})^2)}{\sum_{j=1}^n \exp(-\beta \cdot (\mathbf{I}_j - \bar{I})^2)}, \quad (18)$$

where  $\bar{I}$  is the mean intensity of the  $i$ -th source image,  $\beta$  is a parameter controlling the strength of the weighting function, and  $n$  is the number of source images. By weighing the contribution of each source image to the final fused image based on its similarity to the fused image and its visibility, the MEF-SSIM provides a more accurate and robust measure of the quality of multi-exposure fusion images.

## 4.3. Parameter Settings

All of our experiments were run on one NVIDIA GeForce RTX 4090 GPU. The optimizer chose ADAM, with a learning rate of  $2e-4$ . The weight of the loss function is set to  $\lambda_1 = 1.0$  and  $\lambda_2 = 1.0$ . The level of the Laplacian pyramid decomposition is  $N = 4$ . Before training, we process the data set and scale the image to  $512 \times 512$  size before inputting it into the network for training. Guided by the JPEG compression methodology, we empirically establish a patch size of  $8 \times 8$  for both the weight matrix estimation and the computation of self-attention, ensuring consistency and uniformity across our calculations.

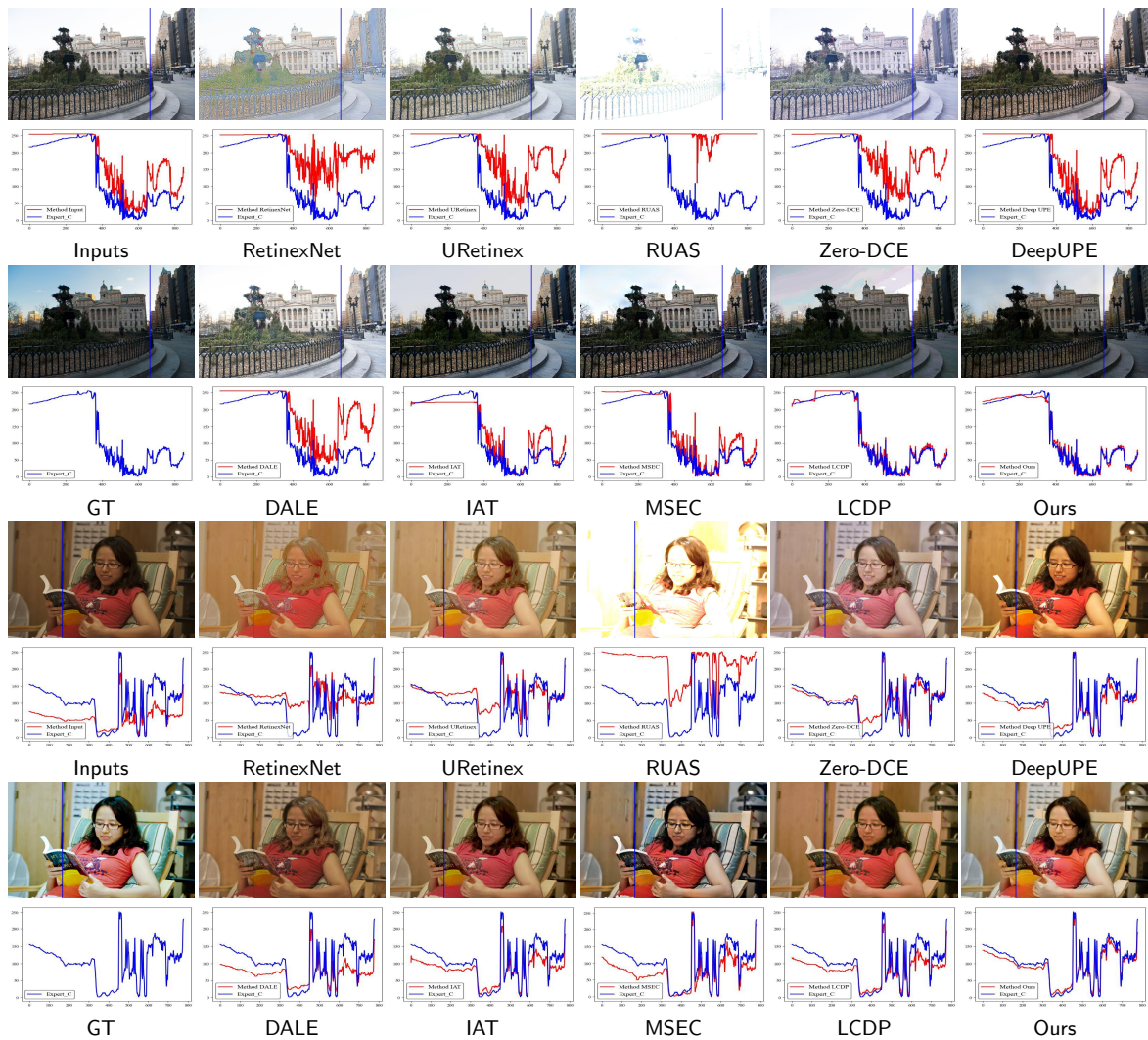


Fig. 4: Qualitative Comparison of exposure correction performance on MSEC dataset.

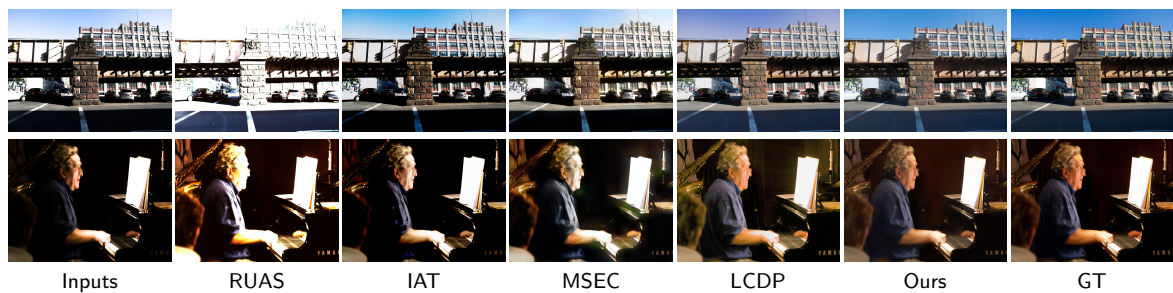


Fig. 5: Qualitative Comparison of exposure correction performance on LCDP dataset.

#### 4.4. Comparison on Exposure Correction

Our proposed method undergoes evaluation on two classic large-scale datasets, with the selection of cutting-edge techniques for comparison. For the MSEC dataset, we compare with methods including: WVM [66], LIME [20], HDR CNN [67], DPED [68], DPE [26], RetinexNet [28], Deep UPE [1], Zero-DCE [2], RUAS [69], URetinex [70], DALE [71], IAT [30], MSEC [3], and LCDP [4]. For the LCDP

dataset, we compare with methods including: Zero-DCE [2], HE [16], RetinexNet [28], CLAME [17], LIME [20], MSEC [3], IAT [30], Deep UPE [1], HDRnet [72], and LCDP [4].

##### 4.4.1. Qualitative Comparison

Figure 4 and Figure 5 respectively display qualitative comparisons on the MSEC and LCDP datasets. Upon an



**Table 1**

Quantitative comparison of exposure correction performance on MSEC dataset.

Method	Expert A		Expert B		Expert C		Expert D		Expert E		Avg	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
WVM	14.488	0.788	15.803	0.699	15.117	0.678	15.863	0.693	16.469	0.704	15.548	0.688
LIME	11.154	0.591	11.828	0.610	11.517	0.607	12.638	0.628	13.613	0.653	12.150	0.618
HDR CNN w/PS	15.812	0.667	16.970	0.699	16.428	0.681	17.301	0.687	18.650	0.702	17.032	0.687
DPED (iPhone)	15.134	0.609	16.505	0.636	15.907	0.622	16.571	0.627	17.251	0.649	16.274	0.629
DPED (BlackBerry)	16.910	0.642	18.649	0.713	17.606	0.653	18.070	0.679	18.217	0.668	17.890	0.671
DPE (HDR)	15.690	0.614	16.548	0.626	16.305	0.626	16.147	0.615	16.341	0.633	16.206	0.623
DPE (S-FiveK)	16.933	0.678	17.701	0.668	17.741	0.696	17.572	0.674	17.601	0.670	17.510	0.677
RetinexNet	10.759	0.585	11.613	0.596	11.135	0.605	11.987	0.615	12.671	0.636	11.633	0.607
Deep UPE	13.161	0.610	13.901	0.642	13.689	0.632	14.806	0.649	15.678	0.667	14.247	0.640
Zero-DCE	11.643	0.536	12.555	0.539	12.058	0.544	12.964	0.548	13.769	0.580	12.597	0.549
RUAS	10.166	0.391	10.522	0.440	9.356	0.411	11.013	0.441	11.574	0.466	10.526	0.430
URetinex	11.420	0.632	12.230	0.700	11.818	0.672	13.078	0.701	14.066	0.735	12.522	0.688
DALE	13.294	0.691	14.324	0.757	13.734	0.722	14.256	0.743	14.511	0.763	14.024	0.735
IAT(local)	16.610	0.750	17.520	0.822	16.950	0.780	17.020	0.780	16.430	0.789	16.910	0.783
MSEC	19.158	0.746	20.096	0.734	20.205	0.769	18.975	0.719	18.983	0.727	19.483	0.739
IAT	19.900	0.817	21.650	0.867	21.230	0.850	19.860	0.844	19.340	0.840	20.340	0.844
LCDPNet	20.574	0.809	21.804	0.865	22.295	0.855	20.108	0.824	19.281	0.822	20.812	0.835
<b>Ours</b>	<b>20.795</b>	<b>0.821</b>	<b>21.902</b>	<b>0.874</b>	<b>22.812</b>	<b>0.859</b>	<b>20.113</b>	<b>0.837</b>	<b>19.979</b>	<b>0.836</b>	<b>21.120</b>	<b>0.845</b>

**Table 2**

Quantitative comparison of exposure correction performance on LCDP dataset.

Method	ZeroDCE	HE	RetinexNet	CIAHE	LIME	MSEC	IAT	DeepUPE	HDRnet	LCDP	<b>Ours</b>
PSNR	12.587	15.975	16.201	16.327	17.335	17.066	17.842	20.970	21.834	23.239	<b>23.415</b>
SSIM	0.653	0.684	0.631	0.642	0.686	0.642	0.684	0.818	0.818	0.842	<b>0.851</b>

overall observation, images corrected by our method demonstrate the closest details and colors when compared to reference images. In addition, we verify the results through an intensity signal analysis. For underexposed inputs, Zero-DCE and Deep UPE achieve an overall similarity with the reference image signals, despite local discrepancies. Other methods reveal a large overall deviation from the reference images due to failed global exposure adjustment. For overexposed inputs, IAT, MSEC, and LCDP effectively reduce exposure values but struggle to restore the lost colors, leading to considerable artifact generation. Other methods fail in correctly correcting overexposed images. Our proposed method exhibits the highest accuracy in pixel intensity, closely aligning with the ground truth values.

#### 4.4.2. Quantitative Comparison

We employ PSNR and SSIM as the metrics to measure the quality of corrected images. As Table 1 shows, for the MSEC dataset, we compare our results with those from five expert photographers. This comparison method considers the variations in camera-based rendering settings, where professionals might render the same image differently. Therefore, our paper evaluates the proposed method against five expert-rendered images, all of which represent satisfactory exposure reference images. Our method achieves the highest scores on both PSNR and SSIM across all five expert reference sets. These results indicate that the proposed

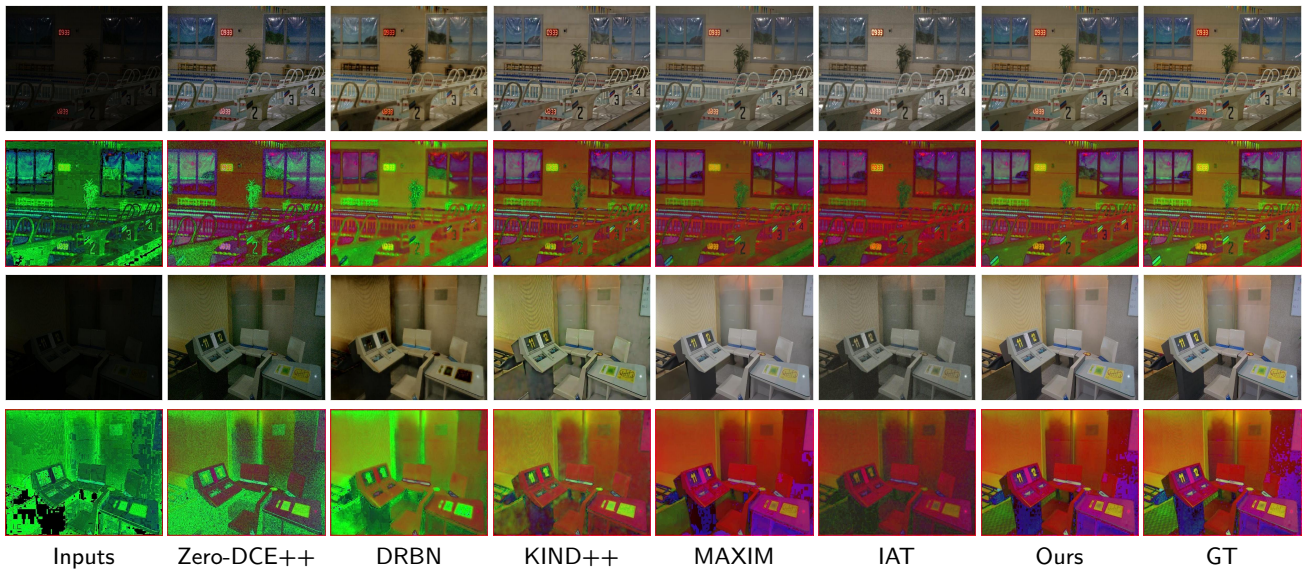
method effectively corrects both overexposed and underexposed input images, delivering high-quality results with accurate pixel intensity distributions and true-to-life textures. Table 2 presents the results for the LCDP dataset, revealing that our proposed method similarly achieves optimal results. This performance demonstrates that our method not only corrects images with global exposure errors but also handles scenarios with different types of exposure errors in different regions. The superior results across both datasets powerfully exhibit the excellent performance of our method in dealing with exposure-error datasets.

#### 4.5. Comparison on Low-light Enhancement

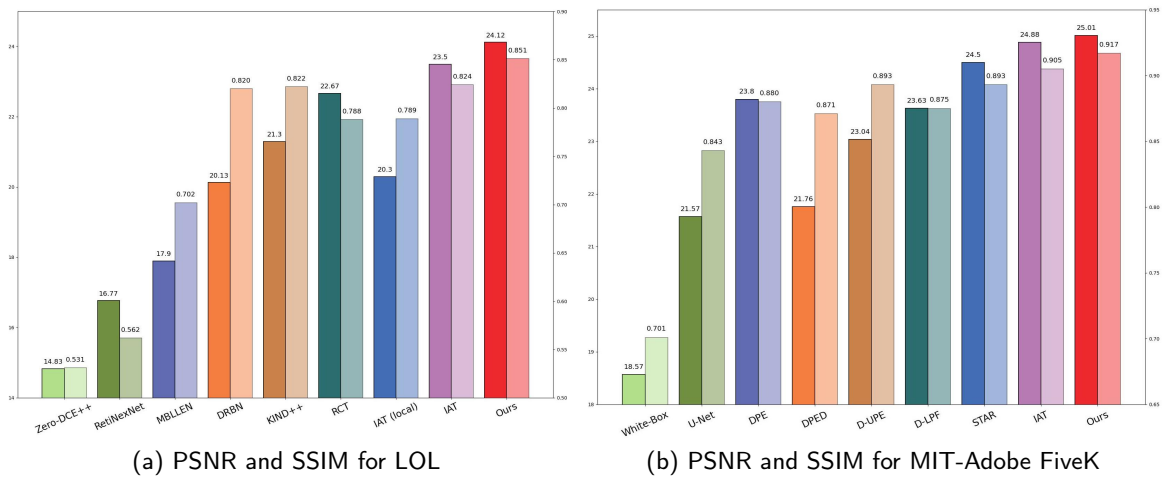
To verify the single exposure correction capability of our method, we not only test its exposure correction performance but also its performance in low-light enhancement tasks. For the LOL dataset [28], we compare our method with several state-of-the-art methods including Zero-DCE++ [2], RetinexNet [28], MBLLEN [73], DRBN [74], KIND++ [75], RCT [76], and IAT [30]. For the MIT-Adobe FiveK dataset [77], we compare our method with White-Box [78], U-Net [79], DPE [26], DPED [68], D-UPE [1], D-LPF [80], STAR [81], and IAT [30].

##### 4.5.1. Qualitative Comparison

A key part of our comparison was a detailed qualitative analysis. To better discern the differences between our method and others, we converted all image enhancement



**Fig. 6:** Qualitative comparison of low-light enhancement performance on LOL dataset. The RGB image is converted to HSV color mode at the bottom, so that it is easier to compare the details.



**Fig. 7:** Quantitative comparison of low-light enhancement performance on LOL and MIT-Adobe FiveK dataset.

results to the HSV color space. This transformation made the disparities in performance more pronounced and visually understandable. The color mappings of the results are presented in Figure 6. In this representation, more similar colors after the mapping operation suggest more consistent brightness and color across corresponding areas. Through this process, we found that popular methods such as Zero-DCE++, DRBN, and KIND++ failed to holistically enhance the image, resulting in an overall darker image. Other methods we compared showed larger color discrepancies in certain areas when compared to the ground truth (GT), indicating an inherent limitation in capturing global dependencies, thus failing to enhance challenging, less noticeable dark parts. In stark contrast, our method showcased its ability to enhance both global and local brightness to suitable levels, producing images that not only exhibit accurate brightness levels but also preserve the best details and colors. We have

shown these successful enhancements in various sample sets - the diving platform in the first set, the control panel in the second set, and the box in the third set - attesting to the robustness of our method across different scenarios.

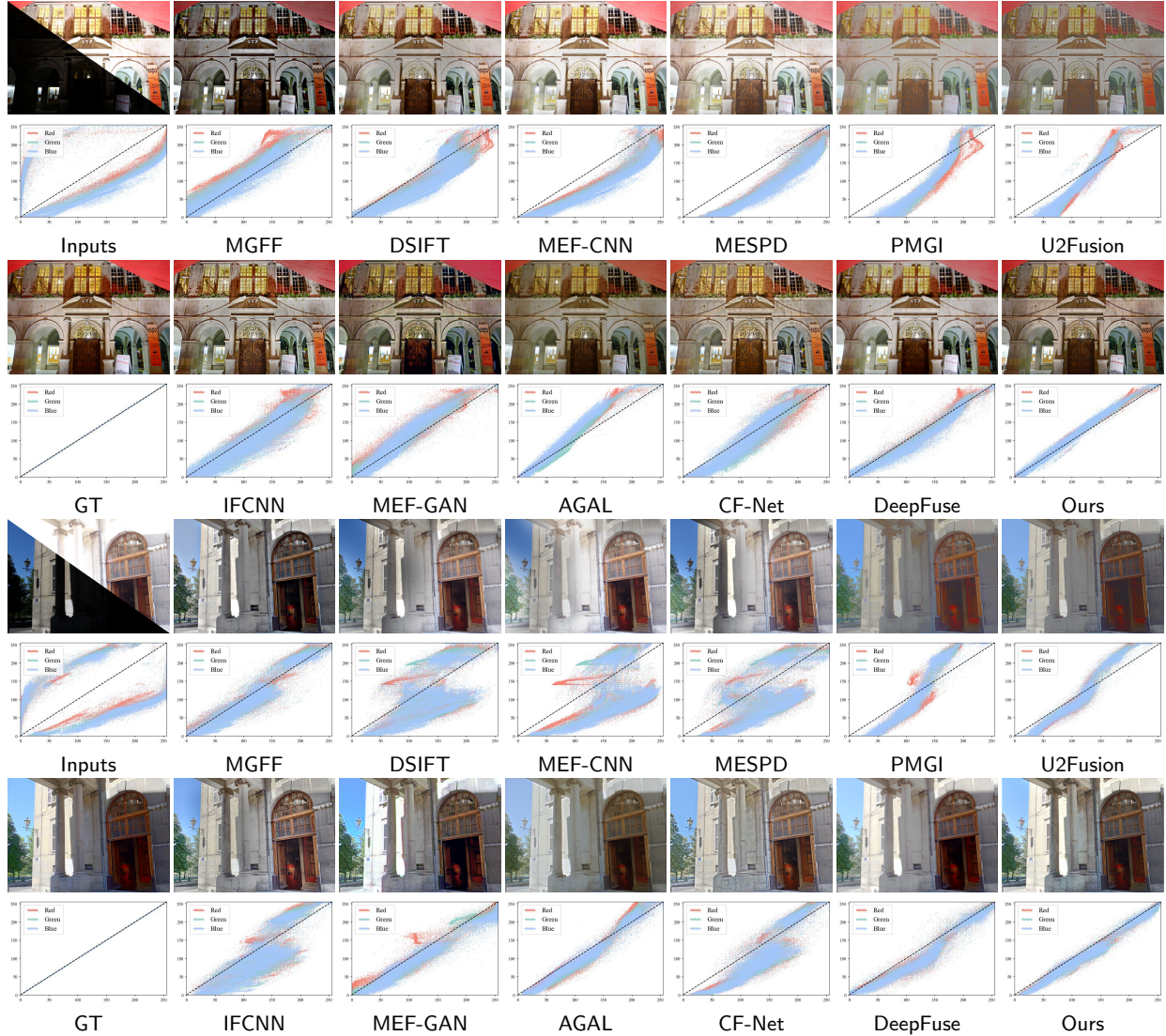
#### 4.5.2. Quantitative Comparison

We visualize the PSNR and SSIM test results of the two datasets using a dual bar chart, as shown in Figure 7. On the LOL dataset, our method was unmatched, achieving the highest scores in both PSNR (24.12) and SSIM (0.851), surpassing all other methods in the test. In fact, our method improved upon the second-best PSNR score (IAT's 23.50) by a significant margin of 2.6% and topped the second-best SSIM score (IAT's 0.824) by an impressive 3.3%. The same superiority of our method was observed on the MIT-Adobe FiveK dataset, where our method again delivered the best performance, achieving a PSNR score of 25.01 and an

**Table 3**

Quantitative comparison of multi-exposure fusion performance on SICE dataset.

Metric	MGFF	PMGI	MEFCNN	MEFCL	DeepFuse	DSIFT	U2Fusion	IFCNN	MEFGAN	AGAL	CFNET	DPEMEF	Ours
PSNR	19.19	17.42	14.19	19.32	17.58	15.38	17.67	19.13	19.71	19.91	20.35	19.23	<b>21.45</b>
SSIM	0.894	0.868	0.752	0.901	0.883	0.813	0.863	0.894	0.902	0.919	0.908	0.904	<b>0.941</b>
MEF-SSIM	0.820	0.899	0.875	0.908	0.843	0.848	0.897	0.896	0.819	0.868	0.904	0.916	<b>0.926</b>

**Fig. 8:** Qualitative comparison of multi-exposure fusion performance on SICE dataset. The signal contrast below the image is derived from the per-channel RGB mapping from the fused image to the GT image.

SSIM score of 0.917. Compared to the second-best method (IAT), our method enhanced the PSNR and SSIM scores by 0.5% and 1.3%, respectively. These results highlight the robustness and consistency of our method in enhancing images with minimal loss of structural and textural information. The comprehensive experimental comparison we have conducted showcases the outstanding performance of our method in low-light enhancement.

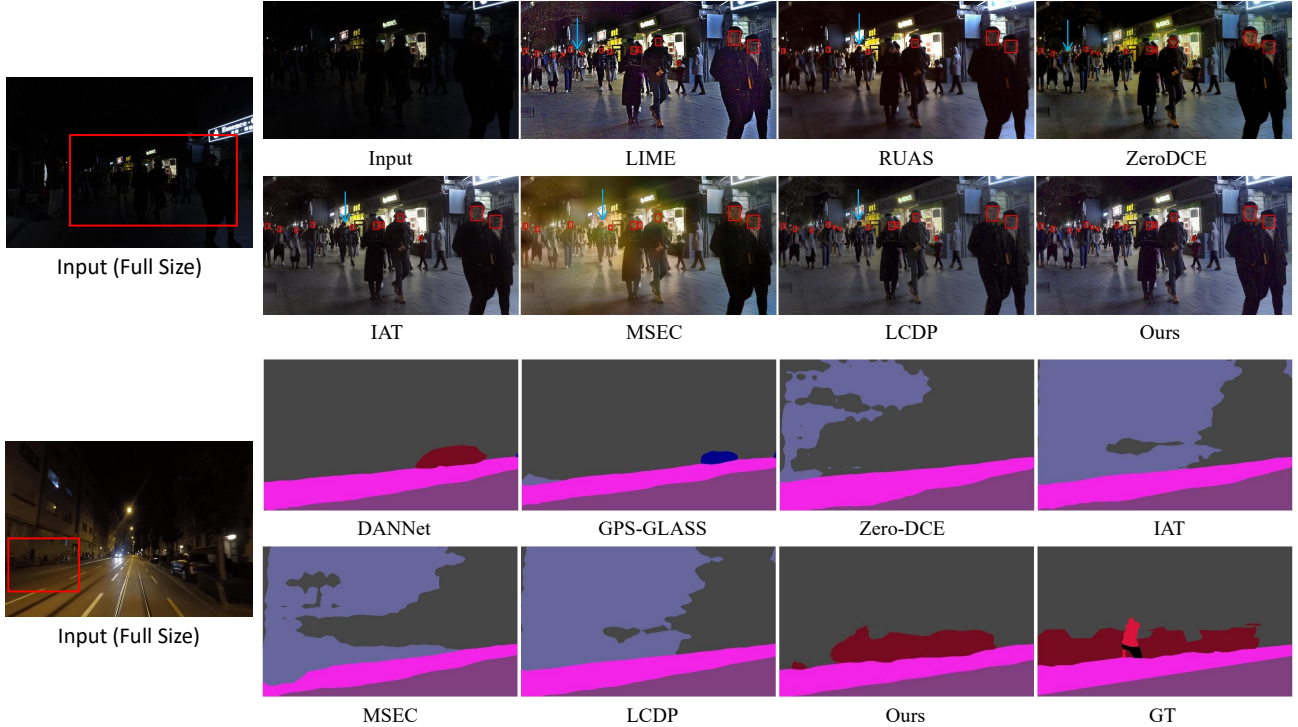
#### 4.6. Comparison on Multi-Exposure Fusion

Our method is also used in the processing of exposure fusion. It is subjected to a comparative study against eleven leading-edge algorithms. This includes a trio of classic algorithms, specifically, MGFF[6], DSIFT[7], MEF-CNN[82], and a group of eight advanced algorithms built upon the foundations of deep learning, such as, DeepFuse[8], PMGI[9], U2Fusion[83], MEF-GAN[84], AGAL[52], CF-NET[85], and DPE-MEF[10].

**Table 4**

Performance comparisons on high-level vision tasks. We retrain the detector/segmentator in all cases containing the enhancer.

Task	Dark Face Detector			Enhancer + Detector (Finetune)						
Method	HLA	REG	MEAT	LIME	ZeroDCE	MSEC	RUAS	LCDP	IAT	<b>Ours</b>
mAP	0.607	0.514	0.526	0.644	0.665	0.659	0.642	0.654	0.663	<b>0.677</b>
Task	Nighttime Semantic Segmentator			Enhancer + Segmentator (Finetune)						
Method	DANNet	CIC	GPS-GLASS	LIME	ZeroDCE	MSEC	RUAS	LCDP	IAT	<b>Ours</b>
mIoU	0.398	0.264	0.380	0.447	0.452	0.449	0.448	0.455	0.456	<b>0.468</b>

**Fig. 9:** Qualitative comparison on high-level tasks.

#### 4.6.1. Qualitative Comparison

The qualitative comparison with other methods is depicted in Figure 8. To discern the difference between each fused method and the Ground Truth (GT) image more clearly, we map each pixel in the RGB channel of each image to the corresponding pixel in the RGB channel of the GT image and depict it beneath each image group. The closer the RGB mapping curve is to the diagonal, the closer each fused image is to the details and colors of the GT image. The images fused by MGFF, DSIFT, MEF-CNN, MESPD, PMGI, and U2Fusion exhibit significant exposure misalignment and ghosting, as indicated by the large deviation of the mapping curve from the GT. Other comparison methods tend to be linear overall, but suffer from localized detail distortion, resulting in significant bulges in the mapping curve. In contrast, our method can achieve the best colors and details after fusion, and the RGB mapping curve is the smoothest and closest to the diagonal.

#### 4.6.2. Quantitative Comparison

Table 3 presents the comparison of our method with others in terms of three key metrics: PSNR, SSIM, and MEF-SSIM. As can be observed, our method significantly outperforms all existing methods specifically designed for multi-exposure fusion. This highlights the powerful generalization ability of our proposed method, which can not only correct the direct input exposure errors, but also reasonably correct the exposure after the fusion of the blocks to generate the image.

#### 4.7. Extending to High-Level Tasks

To demonstrate the generalizability of our method, we apply it to relevant high-level vision tasks, including low-light face detection and low-light semantic segmentation tasks. To thoroughly evaluate its performance, we contrast it not only with some correction methods but also consider specific detection methods including HLA [86], REG [87], MAET [88], and segmentation methods including DANNet [70], CIC [89], GPS-GLASS [90].



**Fig. 10:** Visualization of the exposure compensation. The compensation map below the image is obtained by performing absolute subtraction between the corrected image and the GT image.

#### 4.7.1. Qualitative Comparison

We apply our proposed method as an enhancer for low-light images, with the enhanced images input into the baseline network for low-light face detection and low-light semantic segmentation. The comparison results with other methods are shown in Figure 9. For the results of low-light face detection, images enhanced by other methods noticeably suffer from detail distortion and artifacts. These issues severely impede the subsequent network’s detection performance, causing the network to struggle in detecting smaller faces at distance and side faces in close proximity. In contrast, the enhanced images provided by our method achieved the best detection results, effectively identifying these challenging cases. For the results of low-light semantic segmentation, the images enhanced by other methods have limited capabilities for enhancing seriously dark parts, making small-sized objects within difficult to recognize and segment effectively. Our method can effectively restore extremely dark areas for the subsequent network to perform better segmentation.

#### 4.7.2. Quantitative Comparison

As shown in Table 4, our method outperforms existing low-light detection and segmentation methods, which do not directly enhance images but proceed with direct detection and segmentation, forming a cascade training network. As the built-in enhancement module is coupled with the detector, it’s challenging to further improve. In contrast, the method of enhancing first and detecting later realizes information decoupling, allowing each part to be optimized separately. Under the same baseline network, we outperform all other enhancement methods in both detection and segmentation aspects.

### 4.8. Ablation Study

#### 4.8.1. Study on Exposure Compensation

To delve deeper into the specifics of the exposure gain performed by our proposed method when processing input images, we carried out an extensive analysis. Exposure gain, in our case, is accomplished by the absolute difference between the corrected and input images. The visualization of exposure gains for every comparative method is presented in Figure 10.

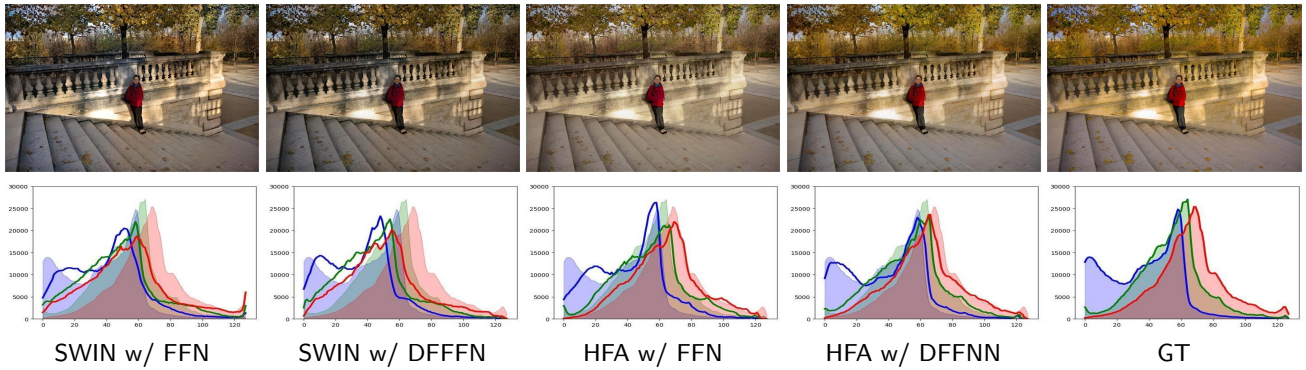


Fig. 11: Visualization of the proposed module ablation experiment.

Table 5

Ablation studies on the proposed Holistic Frequency Attention and Dynamic Frequency FeedForward Network.

Options	Attention		Feed-Forward Network		Metrics	
	HF Attention	SWIN Attention	DF FFN	FFN	PSNR	SSIM
SWIN w/ FFN	✗	✓	✗	✓	22.079	0.821
SWIN w/ DFFFN	✗	✓	✓	✗	22.214	0.827
HFA w/ FFN	✓	✗	✗	✓	22.571	0.832
HFA w/ DFFNN	✓	✗	✓	✗	22.812	0.859

As seen from the figure, our method demonstrates superior performance in achieving optimal exposure gain. It not only accurately identifies every region requiring gain but also applies appropriate enhancement to these areas. Moreover, our approach is capable of increasing details and colors in areas that have already suffered significant distortion. In comparison, the other methods either fail to provide correct exposure gain, or they are unable to effectively enhance details and colors. This comparative analysis further underlines the effectiveness and precision of our proposed method in terms of exposure gain, especially when dealing with areas that exhibit distortion in details and colors.

#### 4.8.2. Study on Holistic Frequency Attention

The spectral domain attention that we propose has significantly reduced the computational complexity. While the original attention's complexity is  $O(N^2)$ , which is unacceptable for larger input sizes, our method scales at  $O(N * \log(N))$ , primarily due to the properties of the fast Fourier transform. As illustrated in Table 6, the complexity does not increase with the size of the window, unlike window-based methods where the complexity gradually shifts from  $O(N)$  to  $O(N^2)$  as the window size increases. Our approach remains insensitive to the window size, and neither the FLOPs nor the GPU memory usage increase with the window size.

On the other hand, our approach replaces the spatial matrix multiplication with the multiplication in the frequency domain. This explicit use of frequency domain filtering exhibits superior performance in image restoration tasks when compared to the spatial domain. As Table 5 shows, replacing Holistic Frequency Attention with Swin Attention in both DF FFN and FFN leads to a significant drop in PSNR and SSIM. The primary reason is that while the shifted window

Table 6

The computational complexity and test-time costs of frequency domain methods and spatial domain methods.

Window Size	Swin Transformer Block		Ours HDF Block	
	FLOPs	GPU Memory	FLOPs	GPU Memory
8 × 8	37.8G	6.8GB	31.3G	5.7GB
32 × 32	41.7G	11.5GB	31.2G	5.6GB
64 × 64	Out of memory	Out of memory	30.9G	5.3GB
128 × 128	Out of memory	Out of memory	30.8G	5.2GB
512 × 512	Out of memory	Out of memory	29.9G	5.2GB

partitioning method reduces the computational cost, it does not fully leverage the useful information between different windows. This verifies that attention based on frequency domain estimation outperforms window-based attention in exposure correction. Furthermore, we visualize the ablation results and show them in Figure 11. It can be observed that the removal of HFA results in exposure imbalance and detail distortion in images.

#### 4.8.3. Study on Dynamic Frequency FFN

We further verified the effectiveness of Dynamic Frequency FFN. As shown in Table 5, regardless of whether it is under the condition of Holistic Frequency Attention or Swin Attention, removing DF FFN leads to a decline in PSNR and SSIM, proving that the frequency-domain feed-forward network is effective. In addition, the visual comparisons in Figure 1 also demonstrate that the removal of DF FFN results in incorrect adjustment of image exposure conditions. This further confirms the importance of using a learnable matrix instead of a convolutional kernel in image restoration tasks.

**Table 7**

Exploring the effect of Laplacian Pyramid Decomposition. R1 means replacing the U-Net Restorer with HDFformer in stage 1 of the pipeline.

R-1	R-2	R-3	R-4	Lpls Decomposition	PSNR $\uparrow$	SSIM $\uparrow$
-	-	-	-	$\times$	22.576	0.841
$\times$	$\times$	$\times$	$\times$	$\checkmark$	20.205	0.769
$\checkmark$	$\times$	$\times$	$\times$	$\checkmark$	21.682	0.813
$\checkmark$	$\checkmark$	$\times$	$\times$	$\checkmark$	22.198	0.829
$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	<b>22.812</b>	<b>0.859</b>
$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	22.809	0.857

#### 4.8.4. Study on Laplacian Pyramid Decomposition

To further elucidate the efficacy of our proposed pipeline, we conducted an ablation study. Our pipeline is primarily divided into four stages. For baseline comparison, we initialized each stage with a standard U-Net architecture. The stages are then incrementally swapped out for HDFformer, our proposed module, in a hierarchical manner.

As shown in Table 7, a progressive improvement in both PSNR and SSIM scores is observed as U-Nets were incrementally replaced by HDFformer modules. However, a saturation point is reached when all U-Nets were entirely substituted by HDFformer, showing a slight decline in the performance metrics. This subtle degradation can be attributed to the overfitting tendencies of pure transformer architectures. Our findings corroborate that a balanced blend of convolutional layers and transformers, specifically the combination denoted as R1-R2-R3, leads to optimized performance. Interestingly, when the Laplacian Pyramid Decomposition is removed, the performance remained commendable, though not as optimal as the full-fledged version. This highlights the robustness of HDFformer while also indicating that the Laplacian Pyramid Decomposition contributes to the superiority of the full model.

## 5. Conclusions

In this paper, we have proposed a spectral-domain attention and feed-forward network. We have validated their potential in both low-level and high-level tasks related to exposure, including exposure correction, low-light enhancement, multi-exposure fusion, low-light face detection, and low-light semantic segmentation. These tasks, which require sophisticated manipulation and understanding of image exposure, benefit significantly from our proposed spectral domain operations. Unlike traditional attention and feed-forward operations that operate in the spatial domain, our methods leverage the unique properties of the frequency domain to achieve superior performance with lower computational overhead.

## 6. Acknowledgements

This work is partially supported by China Postdoctoral Science Foundation (2023M730741), and the National Natural Science Foundation of China (Nos.62302078)

## References

- [1] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6849–6857, 2019.
- [2] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020.
- [3] Mahmoud Afifi, Konstantinos G Derpanis, Bjorn Ommer, and Michael S Brown. Learning multi-scale photo exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9157–9167, 2021.
- [4] Haoyuan Wang, Ke Xu, and Rynson WH Lau. Local color distributions prior for image enhancement. In *European Conference on Computer Vision*, pages 343–359. Springer, 2022.
- [5] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep fourier-based exposure correction network with spatial-frequency interaction. In *European Conference on Computer Vision*, pages 163–180. Springer, 2022.
- [6] Shutao Li, Xudong Kang, and Jianwen Hu. Image fusion with guided filtering. *IEEE Transactions on Image Processing*, 22(7):2864–2875, 2013.
- [7] Yu Liu and Zengfu Wang. Dense sift for ghost-free multi-exposure fusion. *Journal of Visual Communication and Image Representation*, 31:208–224, 2015.
- [8] K. R. Prabhakar, V. S. Srikanth, and R. V. Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *IEEE International Conference on Computer Vision*, pages 4724–4732, 2017.
- [9] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *AAAI Conference on Artificial Intelligence*, pages 12797–12804, 2020.
- [10] Dong Han, Liang Li, Xiaojie Guo, and Jiayi Ma. Multi-exposure image fusion via deep perceptual enhancement. *Information Fusion*, 79:248–262, 2022.
- [11] Junchao Zhang, Yidong Luo, Junbin Huang, Ying Liu, and Jiayi Ma. Multi-exposure image fusion via perception enhanced structural patch decomposition. *Information Fusion*, page 101895, 2023.
- [12] Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Information Fusion*, page 101870, 2023.
- [13] Xingyu Hu, Junjun Jiang, Xianming Liu, and Jiayi Ma. Zmf: Zero-shot multi-focus image fusion. *Information Fusion*, 92:127–138, 2023.
- [14] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma. Divfusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91:477–493, 2023.
- [15] Mohammad Abdullah-Al-Wadud, Md Hasanul Kabir, M Ali Akber Dewan, and Oksam Chae. A dynamic histogram equalization for image contrast enhancement. *IEEE transactions on consumer electronics*, 53(2):593–600, 2007.
- [16] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.
- [17] Ali M Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38:35–44, 2004.
- [18] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. A new image contrast enhancement algorithm using exposure fusion framework. In *Computer Analysis of Images and Patterns: 17th International Conference, CAIP 2017, Ystad, Sweden, August 22-24,*

- 2017, *Proceedings, Part II 17*, pages 36–46. Springer, 2017.
- [19] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977.
- [20] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016.
- [21] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE Transactions on Image Processing*, 27(6):2828–2841, 2018.
- [22] Xutong Ren, Wenhan Yang, Wen-Huang Cheng, and Jiaying Liu. Lr3m: Robust low-light enhancement via low-rank regularized retinex model. *IEEE Transactions on Image Processing*, 29:5862–5876, 2020.
- [23] Bolun Cai, Xianming Xu, Kailing Guo, Kui Jia, Bin Hu, and Dacheng Tao. A joint intrinsic-extrinsic prior model for retinex. In *Proceedings of the IEEE international conference on computer vision*, pages 4000–4009, 2017.
- [24] Long Ma, Dian Jin, Nan An, Jinyuan Liu, Xin Fan, and Risheng Liu. Bilevel fast scene adaptation for low-light image enhancement. *arXiv preprint arXiv:2306.01343*, 2023.
- [25] Risheng Liu, Xin Fan, Ming Zhu, Minjun Hou, and Zhongxuan Luo. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *IEEE transactions on circuits and systems for video technology*, 30(12):4861–4875, 2020.
- [26] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6306–6314, 2018.
- [27] Feifan Lv, Yu Li, and Feng Lu. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *International Journal of Computer Vision*, 129(7):2175–2193, 2021.
- [28] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.
- [29] Long Ma, Tianjiao Ma, Xinwei Xue, Xin Fan, Zhongxuan Luo, and Risheng Liu. Practical exposure correction: Great truths are always simple. *arXiv preprint arXiv:2212.14245*, 2022.
- [30] Ziteng Cui, Kunchang Li, Lin Gu, Shenghan Su, Peng Gao, Zhengkai Jiang, Yu Qiao, and Tatsuya Harada. Illumination adaptive transformer. *arXiv preprint arXiv:2205.14871*, 2022.
- [31] Xiaojie Guo, Yang Yang, Chaoyue Wang, and Jiayi Ma. Image dehazing via enhancement, restoration, and fusion: A survey. *Information Fusion*, 86:146–170, 2022.
- [32] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022.
- [33] Dong Han, Liang Li, Xiaojie Guo, and Jiayi Ma. Multi-exposure image fusion via deep perceptual enhancement. *Information Fusion*, 79:248–262, 2022.
- [34] Di Wang, Jinyuan Liu, Risheng Liu, and Xin Fan. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. *Information Fusion*, 98:101828, 2023.
- [35] Jinyuan Liu, Guanyao Wu, Junsheng Luan, Zhiying Jiang, Risheng Liu, and Xin Fan. Holoco: Holistic and local contrastive learning network for multi-exposure image fusion. *Information Fusion*, 95:237–249, 2023.
- [36] Jinyuan Liu, Xin Fan, Ji Jiang, Risheng Liu, and Zhongxuan Luo. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):105–119, 2021.
- [37] Risheng Liu, Jinyuan Liu, Zhiying Jiang, Xin Fan, and Zhongxuan Luo. A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. *IEEE Transactions on Image Processing*, 30:1261–1274, 2020.
- [38] Jinyuan Liu, Yuhui Wu, Zhanbo Huang, Risheng Liu, and Xin Fan. Smoa: Searching a modality-oriented architecture for infrared and visible image fusion. *IEEE Signal Processing Letters*, 28:1818–1822, 2021.
- [39] Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1600–1608, 2021.
- [40] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European Conference on Computer Vision*, pages 539–555. Springer, 2022.
- [41] Risheng Liu, Long Ma, Tengyu Ma, Xin Fan, and Zhongxuan Luo. Learning with nested scene modeling and cooperative architecture search for low-light vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5953–5969, 2022.
- [42] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022.
- [43] Long Ma, Risheng Liu, Yiyang Wang, Xin Fan, and Zhongxuan Luo. Low-light image enhancement via self-reinforced retinex projection model. *IEEE Transactions on Multimedia*, 2022.
- [44] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *arXiv preprint arXiv:2211.10960*, 2022.
- [45] Risheng Liu, Xin Fan, Minjun Hou, Zhiying Jiang, Zhongxuan Luo, and Lei Zhang. Learning aggregated transmission propagation networks for haze removal and beyond. *IEEE transactions on neural networks and learning systems*, 30(10):2973–2986, 2018.
- [46] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017.
- [47] Kede Ma, Zhengfang Duanmu, Hanwei Zhu, Yuming Fang, and Zhou Wang. Deep guided learning for fast multi-exposure image fusion. *IEEE Transactions on Image Processing*, 29:2808–2819, 2019.
- [48] Jia-Li Yin, Bo-Hao Chen, Yan-Tsung Peng, and Chung-Chi Tsai. Deep prior guided network for high-quality image fusion. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.
- [49] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Proceedings of the IEEE international conference on computer vision*, pages 4714–4722, 2017.
- [50] Han Xu, Jiayi Ma, and Xiao-Ping Zhang. Mef-gan: Multi-exposure image fusion via generative adversarial networks. *IEEE Transactions on Image Processing*, 29:7203–7216, 2020.
- [51] Sheng-Yeh Chen and Yung-Yu Chuang. Deep exposure fusion with dehazing via homography estimation and attention learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1464–1468. IEEE, 2020.
- [52] Jinyuan Liu, Jingjie Shang, Risheng Liu, and Xin Fan. Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5026–5040, 2022.
- [53] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7254–7263, 2019.
- [54] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9060–9069, 2020.
- [55] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for rgb-d saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3472–3481, 2020.



- [56] Risheng Liu, Zhouchen Lin, Fernando De la Torre, and Zhixun Su. Fixed-rank representation for unsupervised visual learning. In *2012 IEEE conference on computer vision and pattern recognition*, pages 598–605. IEEE, 2012.
- [57] Jianlong Wu, Zhouchen Lin, and Hongbin Zha. Essential tensor learning for multi-view spectral clustering. *IEEE Transactions on Image Processing*, 28(12):5910–5922, 2019.
- [58] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in neural information processing systems*, 34:980–993, 2021.
- [59] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021.
- [60] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020.
- [61] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023.
- [62] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017.
- [63] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
- [64] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020.
- [65] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021.
- [66] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2782–2790, 2016.
- [67] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017.
- [68] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3277–3285, 2017.
- [69] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan 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*, pages 10561–10570, 2021.
- [70] Jun Xu, Yingkun Hou, Dongwei Ren, Li Liu, Fan Zhu, Mengyang Yu, Haoqian Wang, and Ling Shao. Star: A structure and texture aware retinex model. *IEEE Transactions on Image Processing*, 29:5022–5037, 2020.
- [71] Dokyeon Kwon, Guisik Kim, and Junseok Kwon. Dale: Dark region-aware low-light image enhancement. *arXiv preprint arXiv:2008.12493*, 2020.
- [72] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédéric Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.
- [73] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mblen: Low-light image/video enhancement using cnns. In *BMVC*, volume 220, page 4, 2018.
- [74] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying 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*, pages 3063–3072, 2020.
- [75] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *International Journal of Computer Vision*, 129:1013–1037, 2021.
- [76] Hanul Kim, Su-Min Choi, Chang-Su Kim, and Yeong Jun Koh. Representative color transform for image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4459–4468, 2021.
- [77] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédéric Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. IEEE, 2011.
- [78] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics (TOG)*, 37(2):1–17, 2018.
- [79] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [80] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deepplf: Deep local parametric filters for image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12826–12835, 2020.
- [81] Zhaoyang Zhang, Yitong Jiang, Jun Jiang, Xiaogang Wang, Ping Luo, and Jinwei Gu. Star: A structure-aware lightweight transformer for real-time image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4106–4115, 2021.
- [82] Hui Li and Lei Zhang. Multi-exposure fusion with cnn features. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1723–1727. IEEE, 2018.
- [83] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.
- [84] Han Xu, Jiayi Ma, and Xiao-Ping Zhang. Mef-gan: Multi-exposure image fusion via generative adversarial networks. *IEEE Transactions on Image Processing*, 29:7203–7216, 2020.
- [85] Xin Deng, Yutong Zhang, Mai Xu, Shuhang Gu, and Yiping Duan. Deep coupled feedback network for joint exposure fusion and image super-resolution. *IEEE TIP*, 30:3098–3112, 2021.
- [86] Wenjing Wang, Xinhao Wang, Wenhan Yang, and Jiaying Liu. Un-supervised face detection in the dark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1250–1266, 2022.
- [87] Jinxiu Liang, Jingwen Wang, Yuhui Quan, Tianyi Chen, Jiaying Liu, Haibin Ling, and Yong Xu. Recurrent exposure generation for low-light face detection. *IEEE Transactions on Multimedia*, 24:1609–1621, 2021.
- [88] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2553–2562, 2021.
- [89] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4399–4409, 2021.
- [90] Hongjae Lee, Changwoo Han, and Seung-Won Jung. Gps-glass: Learning nighttime semantic segmentation using daytime video and gps data. *arXiv preprint arXiv:2207.13297*, 2022.