Ghost-Free HDR Imaging in Dynamic Scenes via High-Low Frequency Decomposition

Xiang Zhang *, Genggeng Chen, Fan Zhang, Yongzhong Zhang

Posted Date: 15 October 2025

doi: 10.20944/preprints202510.1208.v1

Keywords: High Dynamic Range Imaging; Ghost-Free HDR; High-Low Frequency Decomposition

Article

Ghost-Free HDR Imaging in Dynamic Scenes via High-Low Frequency Decomposition

Xiang Zhang [1,*], Genggeng Chen [1], Fan Zhang [1] and Yongzhong Zhang [2]

[1] College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an, China
[2] China United Network Communications Group Co., Ltd. Shaanxi Branch, Xi'an, China
* Correspondence: zhangxiang@xauat.edu.cn

Abstract

Generating high-quality high dynamic range (HDR) images in dynamic scenes remains a challenging task. Recently, Transformers have been introduced into HDR imaging and have demonstrated superior performance over traditional convolutional neural networks (CNNs) in handling large-scale motion. However, due to the low-pass filtering nature of self-attention, Transformers tend to weaken the capture of high-frequency information, which impairs the recovery of structural details. In addition, their high computational complexity limits practical applications. To address these issues, we propose HL-HDR, a high-low frequency-aware ghost-free HDR reconstruction network for dynamic scenes. By decomposing features into high and low-frequency components, HL-HDR effectively overcomes the limitations of existing Transformer and CNN-based methods. The Frequency Alignment Module (FAM) captures large-scale motion in the low-frequency branch while refining local details in the high-frequency branch. The Frequency Decomposition Processing Block (FDPB) fuses local high-frequency details and global low-frequency context, enabling precise HDR reconstruction. Extensive experiments on five public HDR datasets demonstrate that HL-HDR consistently outperforms state-of-the-art methods in both quantitative metrics and qualitative evaluation. The code is publicly available at https://github.com/chengeng0613/HL-HDR_Plus.

Keywords: high dynamic range imaging; ghost-free HDR; high-low frequency decomposition

1. Introduction

Contemporary digital cameras face inherent limitations in capturing the dynamic range of a real scene due to constraints in their sensor capabilities. In contrast, high dynamic range (HDR) imaging overcomes these limitations by covering a broad range of light intensities, providing a more accurate representation of the real-world scenario. The generation of HDR images encompasses a variety of techniques, with one of the most prevalent methods involving the fusion of Low Dynamic Range (LDR) images captured under diverse exposures to reconstruct the HDR image. However, in multi-exposure HDR reconstruction, motion of objects or camera movement can cause inconsistencies in certain regions or information loss due to overexposure, resulting in ghost-like artifacts. This phenomenon presents a significant challenge in multi-exposure HDR imaging.

To tackle the challenges associated with ghosting in HDR imaging, various methodologies have been developed. Traditional techniques commonly employ methods such as alignment-based methods [1,2], rejection-based methods [3–5], and patch-based methods [6,7] to eliminate or align motion regions in images. However, the efficacy of these methods is largely contingent upon the performance of preprocessing techniques, such as optical flow and motion detection. And when dealing with significant scene motion, the results of these methods typically turn out to be rather unsatisfactory. With the advancement of Deep Neural Networks (DNN), several CNN-based methods[8–13] have been applied in ghost-free HDR imaging. Among them, the "alignment-fusion" paradigm has shown remarkable success, especially in scenarios involving large-scale motion. Moreover, Transformer-based

 doi:10.20944/preprints202510.1208.v1

approaches [14,15], which can capture long-distance dependencies, are introduced as an alternative to CNNs. These methods further enhance HDR imaging performance and are adopted by the current mainstream state-of-the-art methods. However, Transformers still face two major challenges in ghost-free HDR imaging. On one hand, local details and global information are crucial for restoring multi-frame HDR content, while the self-attention mechanism of pure Transformers often exhibits a low-pass filtering effect, reducing the variance of input features and overly smoothing patch tokens. This occurs because self-attention essentially averages features across different patches, suppressing high-frequency information that is vital for distinguishing fine structural details, thereby limiting the Transformer's ability to capture high-frequency local details [14]. On the other hand, HDR images are typically high-resolution, and the computational complexity of self-attention grows quadratically with the spatial dimensions of the input feature map. This results in significant computational and memory overhead in high-resolution scenarios, restricting the practical application and scalability of Transformers in high-resolution HDR imaging tasks.

Inspired by the distinct characteristics of high and low-frequency patterns in images, we propose a frequency-decomposition-based ghost-free HDR image reconstruction network. In both the cross-frame alignment and feature fusion stages, features are decomposed into high and low-frequency components and processed according to their respective characteristics. Since high-frequency components represent local structures while low-frequency components characterize global information, we leverage the low-pass filtering property of average pooling (AvgPool) to decouple features into high-resolution high-frequency components and low-resolution low-frequency components.

Specifically, global motion or long-range dependencies can be effectively represented by low-frequency features without requiring high-resolution feature maps, while high-frequency features focus on fine-grained local structures that need high-resolution maps and are better modeled by local operators. Based on this, we adopt a dual-branch architecture in both stages to balance global information and local details.

In the cross-frame alignment stage, we propose the Frequency Alignment Module (FAM). The low-frequency branch employs a lightweight UNet to learn optical flow and align non-reference frames to the reference frame, efficiently capturing large-scale motion while reducing computational cost. Meanwhile, the high-frequency branch combines convolution and attention to adaptively refine edges and textures, suppressing ghosting and preserving structural consistency.

In the feature fusion stage, we design the Frequency Decomposition Processing Block (FDPB). The high-frequency branch uses a Local Feature Extractor (LFE) to capture details and enhance cross-frame high-frequency information, while the low-frequency branch adopts a Global Feature Extractor (GFE) to model long-range dependencies. To alleviate information loss caused by downsampling, we further introduce a Cross-Scale Fusion Module (CSFM) for effective cross-resolution integration.

By integrating FAM and FDPB, we propose the High–Low Frequency-Aware HDR Network (HL-HDR), which consists of two stages: cross-frame alignment and feature fusion. FAM enables accurate motion modeling and detail preservation, while FDPB hierarchically captures both global and local contexts, leading to high-quality, ghost-free HDR reconstruction.

The main contributions are summarized as follows:

- We propose a novel alignment method, FAM, in which the low-frequency branch captures large-scale motion through optical flow alignment, while the high-frequency branch refines local edges and textures, effectively suppressing ghosting.
- The FDPB module, introduced in our work, addresses low-frequency components by employing a multi-scale feature extraction approach in conjunction with Transformer mechanisms to collectively capture global information. For high-frequency components, we employ small convolutional kernels and densely connected residual links to effectively extract local feature information. This strategic design in our model achieves a harmonious balance between speed and precision.

 doi:10.20944/preprints202510.1208.v1

- A plethora of experiments have substantiated that the proposed methodology, denoted as method HL-HDR, attains state-of-the-art (SOTA) performance in HDR imaging tasks. Furthermore, it yields visually appealing outcomes that align with human perceptual aesthetics.

This work is an extended version of our conference paper[16] presented at the International Joint Conference on Neural Networks (IJCNN 2024). Compared with the conference version, it incorporates a substantial amount of new material. 1)To address the issue of ghosting in moving regions, we optimized the cross-frame alignment stage by designing the FAM module: the low-frequency branch aligns features using optical flow, while the high-frequency branch adaptively refines local edges and textures through convolution and attention mechanisms, effectively suppressing ghost artifacts and maintaining structural consistency. 2)We conducted comparative experiments on additional datasets and against the latest methods, fully demonstrating the advantages of our improved approach. 3)The ablation studies are more detailed and clear, thoroughly verifying and analyzing the contributions of each module.

2. Related Work

Presently, HDR deghosting techniques can be primarily classified into alignment-based methods, rejection-based methods, patch-based methods, and CNN-based methods.

2.1. HDR Deghosting Methods

Alignment-based Method. These methods aim to register all LDR images to a reference image using either rigid or non-rigid algorithms. Bogoni [1] utilized optical flow to estimate motion vectors, while Pece and Kautz [5] computed the Median Threshold Bitmap (MTB) for input images to detect regions of motion. Kang *et al*. [17] transformed the intensities of LDR images into the luminance domain by leveraging exposure time information and computed optical flow to identify corresponding pixels among the LDR images. However, both rigid and non-rigid alignment methods exhibit susceptibility to significant motions, occlusions, and variations in brightness, rendering them prone to errors in complex regions.

Rejection-based Methods. Rejection methods, post the global registration procedure, discern and eliminate motion regions within the input data, followed by the fusion of static regions to reconstruct HDR images. Grosch *et al*.[3] devised an error map by assessing color disparities post alignment, aiming to exclude pixels with mismatches. Pece *et al*.[5] identified regions of motion through the utilization of a median threshold bitmap on input LDR images. Jacobs *et al*.[18] pinpointed areas of misalignment via an analytical approach involving weighted intensity variance analysis. These approaches often yield unsatisfactory HDR outcomes as they incur the loss of valuable information while eliminating pixels.

Patch-based Methods. The patch-based methods, involving patch-wise alignment among exposure images for deghosting, have been explored in the literature. Sen *et al*. [7] introduced a patch-based energy minimization method that simultaneously optimizes alignment and reconstruction. In the work by Hu *et al*. [6], an iterative propagation of intensity and gradient information was conducted using a coarse-to-fine schedule. Ma *et al*. [19] proposed an approach based on structural patch decomposition, which dissects an image patch into signal strength, signal structure, and mean intensity components for the reconstruction of ghost-free images. However, it is noteworthy that these methods lack compensation for saturation and are burdened by elevated computational costs.

CNN-based Methods. Kalantari *et al*. [8] initiated the alignment of images using optical flow and subsequently employed a CNN network for their fusion. Yan *et al*. [10] introduced a spatial attention mechanism based on CNN to mitigate issues related to motion and oversaturated regions. Yan *et al*. [20] formulated a non-local module aimed at expanding the receptive field for comprehensive global merging. In their work, Song *et al*. [21] harnessed the benefits of the Transformer's extensive receptive field to globally recover areas affected by motion. Additionally, HyHDR [22] proposed an innovative patch aggregation module grounded in deep learning, strategically fusing valuable information

 doi:10.20944/preprints202510.1208.v1

from non-reference frames. Despite the significant performance breakthroughs achieved by these methodologies, their outcomes in both dynamic and static areas remain somewhat unsatisfactory.

2.2. Vision Transformer

Transformers have demonstrated remarkable success in natural language processing. The multi-head self-attention mechanism utilized in this context effectively captures long-range correlations among word token embeddings. A recent development, Vision Transformer (ViT) [23], has illustrated that a pure Transformer architecture can be directly applied to sequences of non-overlapping image patches, exhibiting excellent performance in image classification tasks. This showcases the versatility of Transformer models beyond natural language applications, extending their efficacy to the domain of computer vision. CA-ViT, proposed by Liu *et al.*[14], leverages the Transformer's capability for capturing long-range dependencies and extracting global feature information, complementing it with the ability of CNN to extract local information. The collaboration between these features has proven to be highly effective. Building upon the Transformer architecture, Zamir *et al.*[24] introduced improvements by incorporating a locally aware Transformer design. This design enhances the model's perception of local image details by introducing local convolution operations within the Transformer. Our approach is inspired by [14,24], strategically handling local and global information differently based on their inherent characteristics.

3. Method

In a series of LDR images with varying exposure levels, images of the same scene are grouped together. Each group comprises three images: underexposed, normally exposed, and overexposed. Our goal is to fuse the information from these three LDR images to reconstruct an HDR image without ghosting artifacts. In previous research[10,25], the authors utilized a set of three LDR images as input, designating the normally exposed image as the reference frame. Using the input images $\{L_1, L_2, L_3\}$, our model derives the HDR image $\hat{H}$ as follows:

$$\hat{H} = f(L_1, L_2, L_3; \theta), \tag{1}$$

where $f(\cdot)$ represents the HDR imaging function, and θ refers to the network's parameters.

3.1. Overview of the HL-HDR Architecture

As shown in Figure 1, the proposed HL-HDR framework consists of two main stages:

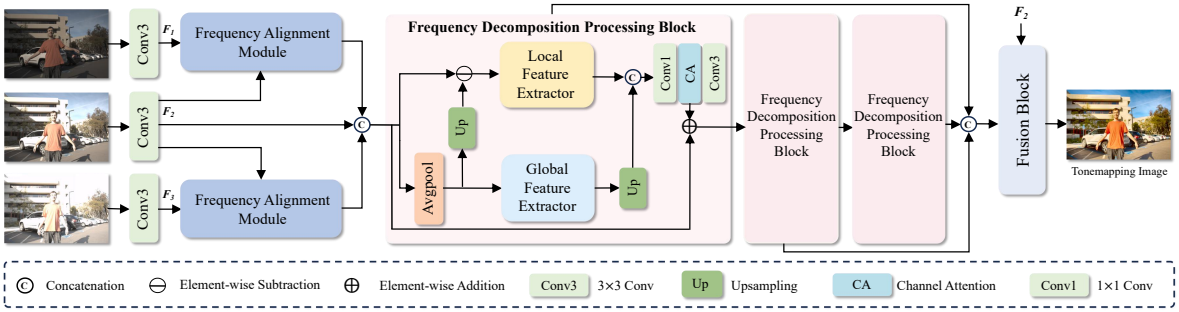


Figure 1. The HL-HDR framework consists of two main stages. First, the cross-frame feature alignment stage uses the Frequency Alignment Module to align overexposed and underexposed images to the reference frame. Second, the feature fusion stage stacks multiple Frequency Decomposition Processing Blocks to extract and integrate features, reconstructing high-quality, ghost-free HDR images.

The first stage is the cross-frame alignment stage, corresponding to the Frequency Alignment Module in the figure. This module takes three images with different exposures as input and extracts shallow features through a shared-weight convolution, producing 64-channel feature maps $\{F_1, F_2, F_3\}$. The

 doi:10.20944/preprints202510.1208.v1

feature map of the normally exposed image, F_2, is used as the reference frame, while the underexposed feature map F_1 and overexposed feature map F_3 are aligned to it.

The second stage is the feature fusion stage, in which multiple Frequency Decomposition Processing Blocks are stacked to extract and integrate the aligned features. Specifically, the features are decomposed into high-frequency and low-frequency components and processed according to their characteristics: the high-frequency branch employs a Local Feature Extractor to capture local details through stacked convolutional blocks and enhances cross-frame high-frequency information via dense connections; the low-frequency branch uses a Global Feature Extractor to model long-range dependencies through multi-layer channel attention. The extracted high and low-frequency features are then fused.

In the final reconstruction stage, the network reduces the number of channels while introducing long-range residual connections, ultimately reconstructing the output as a 3-channel HDR image.

3.2. Frequency Alignment Module

To balance large-scale motion modeling and detail preservation during the alignment stage, we first decompose both the reference frame and the non-reference frames into high-frequency and low-frequency components, which are then independently aligned in separate branches. Specifically, the low-frequency components are obtained by applying average pooling to the original feature maps, while the high-frequency components are derived by subtracting the low-frequency part from the original features. This process can be formally expressed as follows:

$$f_{\mathrm{il}} = \mathrm{Avgpool}(F_{\mathrm{i}}), \tag{2}$$

$$f' = \mathrm{Up}(f_{\mathrm{il}}), f_{\mathrm{ih}} = F_{\mathrm{i}} - f', \tag{3}$$

$$f_{2\mathrm{l}} = \mathrm{Avgpool}(F_2), \tag{4}$$

$$f'' = \mathrm{Up}(f_{2\mathrm{l}}), f_{2\mathrm{h}} = F_2 - f'', \tag{5}$$

where $\mathrm{Avgpool}(\cdot)$ refers to average pooling, $\mathrm{Up}(\cdot)$ stands for bilinear interpolation upsampling. F_{i} refers to the non-reference frame, while F_2 refers to the reference frame.

For the alignment of high-frequency components, the high-frequency features of the reference frame and non-reference frames are concatenated and fed into a convolution-based attention module to generate an attention map, which is subsequently applied to the non-reference frame features. This module is similar to the implicit alignment mechanism in AHDR [10], guiding the network to focus on critical information across different exposures. Under the guidance of attention, the model can adaptively reweight the importance of different regions, thereby refining edges and textures, enhancing local detail restoration, and effectively suppressing ghosting while ensuring structural consistency. Formally, the process can be written as:

$$F_{\mathrm{am}} = \mathrm{AM}(\mathrm{Concat}(f_{\mathrm{ih}}, f_{2\mathrm{h}})), \tag{6}$$

$$F_{\mathrm{h}} = f_{\mathrm{ih}} \cdot F_{\mathrm{am}}, \tag{7}$$

where $\mathrm{AM}(\cdot)$ refers to the attention module, and $(\cdot)$ denotes element-wise multiplication.

For the alignment of low-frequency components, we modify the Encoder-Decoder structure of SAFNet [26] to predict the optical flow field, which is then used to warp the low-frequency components of the non-reference frames, enabling more accurate modeling of large-scale motion. The above operations can be formulated as:

$$F_{\mathrm{am}} = \mathrm{FM}(f_{\mathrm{il}}, f_{2\mathrm{l}}), \tag{8}$$

$$F_{\mathrm{l}} = \mathrm{Warp}(f_{\mathrm{il}}, F_{\mathrm{am}}), \tag{9}$$

 doi:10.20944/preprints202510.1208.v1

where FM$(\cdot)$ refers to the Encoder-Decoder structure shown in Figure 2, and Warp$(\cdot)$ denotes the warping operation.

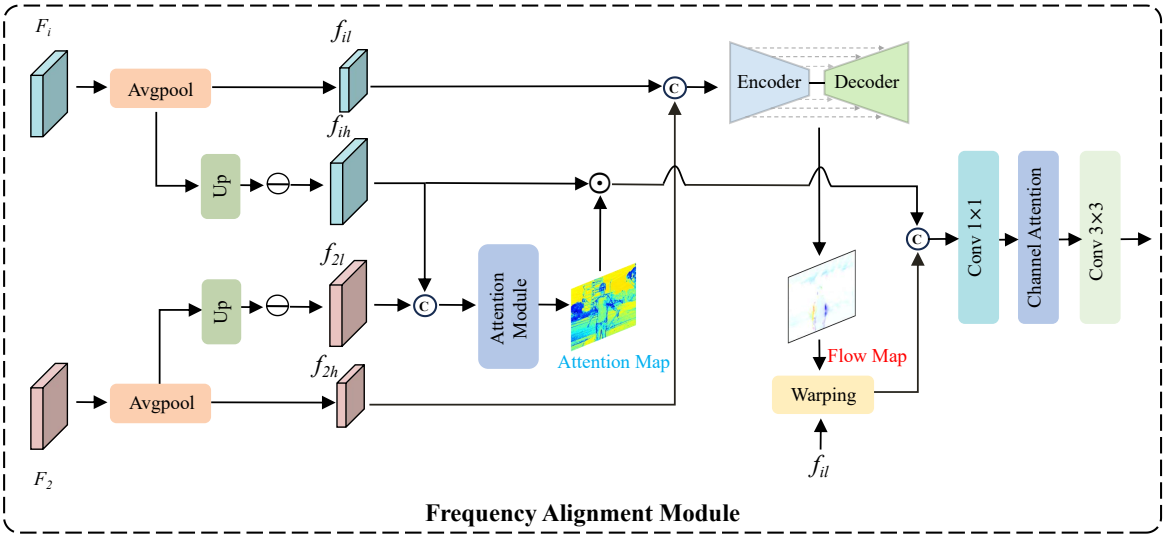


Figure 2. The architecture of the proposed FAM decomposes both the non-reference frames and the reference frame into high-frequency and low-frequency components, which are then aligned separately.

Finally, the aligned low-frequency and high-frequency components are fused, followed by convolution and a channel attention module to restore channel information and further extract features. This process can be formally expressed as follows:

$$F_{\text{out}} = \text{Conv3}(\text{CA}(\text{Conv1}(\text{fusion}(\text{Concat}(F_{\text{l}}, F_{\text{h}}))))), \tag{10}$$

where fusion$(\cdot)$ denotes the operation for integrating features at different scales, CA$(\cdot)$ represents channel attention, Conv1$(\cdot)$ stands for a 1×1 convolution, and Conv3$(\cdot)$ denotes a 3×3 convolution.

3.3. Frequency Decomposition Processing Block

As shown in Figure 1, similar to the alignment stage, we also decompose the feature maps into high-frequency and low-frequency components during the feature fusion stage. By supplementing high-frequency details while extracting the global and background information of the image, this approach enhances local textures and edges, resulting in improved visual quality. This process can be formally expressed as follows:

$$F_{\text{low}} = \text{Avgpool}(F), \tag{11}$$

$$F' = \text{Up}(F_{\text{low}}), F_{\text{high}} = F - F', \tag{12}$$

$$F_{\text{out}} = F + \text{Conv3}(\text{CA}(\text{Conv1}(\text{Concat}(\text{LFE}(F_{\text{high}}), \text{GFE}(F_{\text{low}}))))), \tag{13}$$

where Avgpool$(\cdot)$ refers to average pooling, Up$(\cdot)$ stands for bilinear interpolation upsampling, LFE$(\cdot)$ stands for Local Feature Extractor , GFE$(\cdot)$ stands for Global Feature Extractor, CA$(\cdot)$ denotes channel attention, Conv3$(\cdot)$ stands for 3×3 convolution, and Conv1$(\cdot)$ stands for 1×1 convolution that restores the channel count from 128 to 64.

3.4. Local Feature Extractor

To better recover the detailed information in an image, we process the high-frequency information, which inherently contains a wealth of detail. High-frequency information requires local details, thus the use of convolutions with small kernels allows for a more focused extraction of these details. Furthermore, inspired by the proficiency of standard residual learning in exploring high-frequency information[27–29], we employ dense residual connections when extracting high-frequency informa-

 doi:10.20944/preprints202510.1208.v1

tion. Overall, we utilize six 3×3 convolutions. As depicted in Figure 3, we only display a portion of the residual connections, but in reality, these are dense residual connections. They are not merely connections between adjacent layers, but rather, the output of each layer is merged with the outputs of all preceding layers, enabling each layer to directly access the feature information of all previous layers.

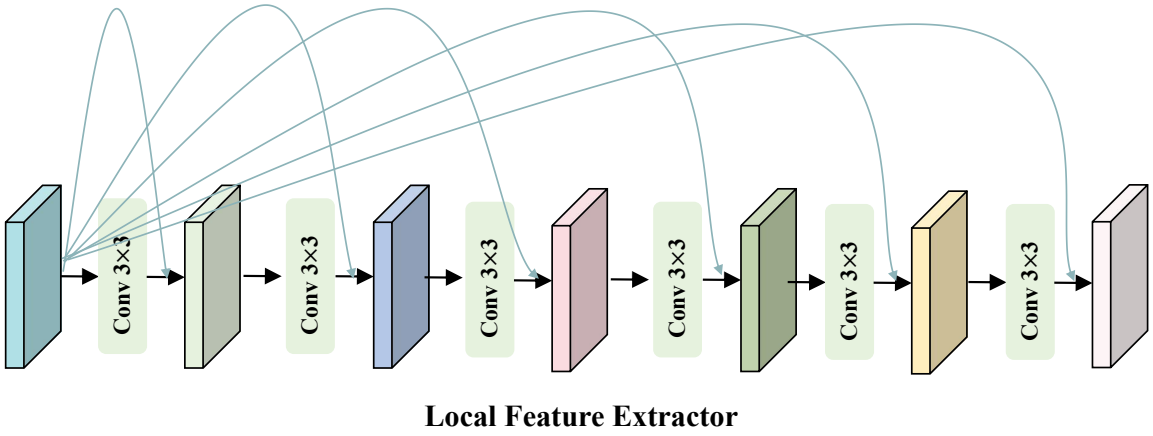


Figure 3. The architecture of the proposed LFE is comprised of a series of standard convolutions and dense residual connections.

3.5. Global Feature Extractor

For low-frequency information, it is necessary to leverage global context to restore the overall structure and background of the image. As shown in Figure 4, although multi-scale feature extraction enables long-range information interactions, some information may be lost during the downsampling process. To address this, we perform feature extraction at each scale to compensate for the information loss caused by downsampling. In each feature extraction layer, we introduce a Channel Transformer Block, which can establish global contextual information and possesses a global receptive field, making it highly suitable for extracting low-frequency features that depend on global information. Furthermore, to compensate for potential information loss when directly upsampling feature maps of different sizes and concatenating them, we employ the CSFM to effectively merge feature maps of varying resolutions.

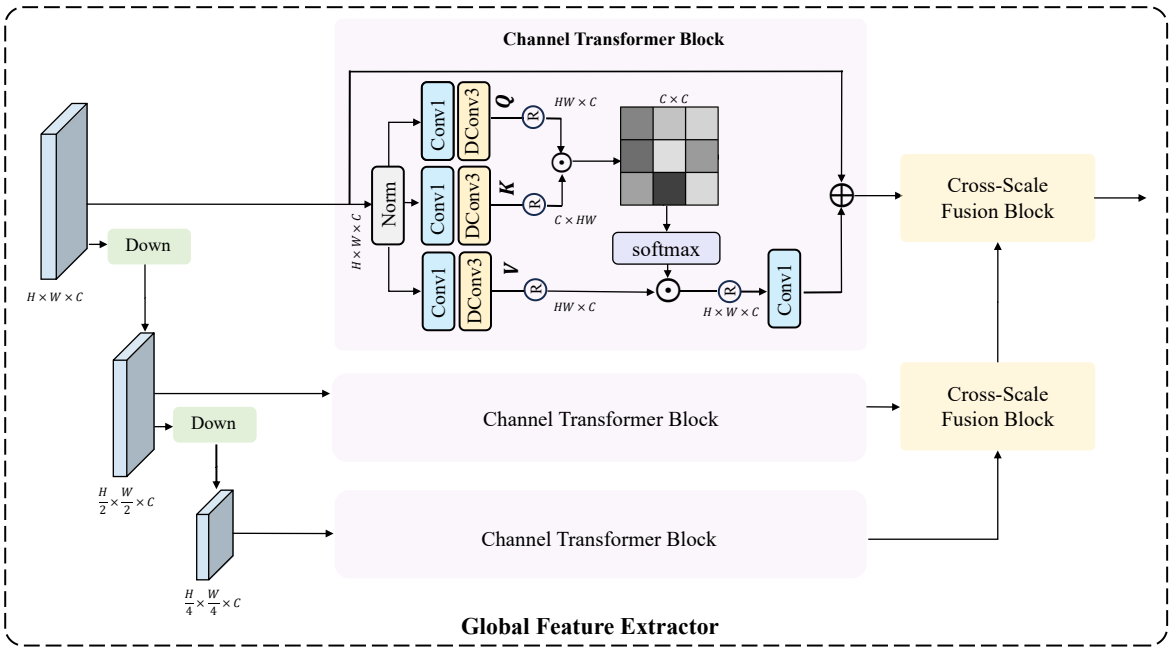


Figure 4. The architecture of the proposed GFE is designed for extracting low-frequency features.

 doi:10.20944/preprints202510.1208.v1

Channel Transformer Block. Inspired by [24], the Transformer structure used here does not employ spatial self-attention, but instead adopts channel-wise self-attention. The input $\mathbf{X} \in R^{H\times W\times C}$ is first layer-normalized to obtain a tensor $\mathbf{Y} \in R^{H\times W\times C}$. Then, 1×1 convolutions are applied to aggregate pixel-wise cross-channel context, followed by 3×3 depth-wise convolutions to encode channel-wise spatial context. This process generates the *query* (**Q**), *key* (**K**), and *value* (**V**), which can be expressed mathematically as:

$$\mathbf{Q} = W_p^Q W_d^Q Y, \mathbf{K} = W_p^K W_d^K Y, \mathbf{V} = W_p^V W_d^V Y, \tag{14}$$

where $W_p(\cdot)$ represents the 1×1 point-wise convolution and $W_d(\cdot)$ represents the 3×3 depth-wise convolution.

Then reshape Q into $R^{HW\times C}$, reshape K into $R^{C\times HW}$. After this transformation, matrix multiplication can be performed, followed by a softmax operation to obtain an attention map $\mathbf{A} \in R^{C\times C}$. Reshape V into $R^{HW\times C}$, allowing for matrix multiplication with A. The resulting output is reshaped into $R^{H\times W\times C}$, and finally, a residual connection is added by summing the initial feature map with the obtained feature map. The specific process is illustrated as follows:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{V} \cdot Softmax(\mathbf{Q} \cdot \mathbf{K}/\alpha), \tag{15}$$

$$\hat{\mathbf{X}} = Attention(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) + \mathbf{X}, \tag{16}$$

where α is a learnable scaling parameter, $\mathbf{X}$ refers to the initial input feature map, and $\hat{\mathbf{X}}$ refers to the final result.

Next, we utilize 1×1 Convolution to aggregate information from different channels and employ 3×3 depth-wise Convolution to aggregate information from spatially neighboring pixels. Additionally, we incorporate a gating mechanism to enhance information encoding. Finally, a long-range residual connection is added, summing the initial feature map with the feature map obtained at this stage.

Cross-Scale Fusion Module. Due to the differing spatial resolutions of features at various scales, they are typically adjusted to a unified resolution via downsampling or upsampling for feature fusion. However, such operations may result in the loss of important structural details, thereby affecting the final image restoration. Inspired by the capability of wavelet transforms to model image scale information [30], we employ wavelet transformation to achieve cross-scale feature fusion. As shown in Figure 5, we use Discrete Wavelet Transform (DWT) to decompose the feature map $F_\mathrm{b} \in R^{H\times W\times C}$ into $(HH, HL, LH, LL) \in R^{\frac{H}{2}\times\frac{W}{2}\times\frac{C}{2}}$, where each feature map has half the width and height of the original feature map, while the number of channels remains unchanged. LL represents the low-frequency information, which is concatenated with the small-scale feature map $F_\mathrm{s} \in R^{\frac{H}{2}\times\frac{W}{2}\times\frac{C}{2}}$, followed by a 1×1 convolution to reduce the number of channels from 128 to 64, and then processed through a ResBlock for further feature extraction. HH, HL, LH represent high-frequency information; after concatenation, a 1×1 convolution reduces the channels to 64, followed by a ResBlock for feature extraction, and finally another 1×1 convolution restores the channel number to 192. The use of two 1×1 convolutions aims to reduce both parameter count and computational cost, as directly extracting information from a 192-channel feature map would be computationally expensive. Finally, the extracted high and low-frequency information undergoes Inverse Discrete Wavelet Transform (IDWT) to obtain the fused final features. This process can be formally expressed as follows:

$$HH, HL, LH, LL = \mathrm{DWT}(F_\mathrm{b}), \tag{17}$$

$$f_\mathrm{b} = \mathrm{Conv1}(\mathrm{Res}(\mathrm{Conv1}(\mathrm{Concat}(HL, LH, LL)))), \tag{18}$$

$$f_\mathrm{s} = \mathrm{Conv1}(\mathrm{Res}(\mathrm{Concat}(F_\mathrm{s}, LL))), \tag{19}$$

$$F = \mathrm{IDWT}(f_\mathrm{b}, f_\mathrm{s}), \tag{20}$$

where $Res(\cdot)$ represents the residual block.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2025 doi:10.20944/preprints202510.1208.v1

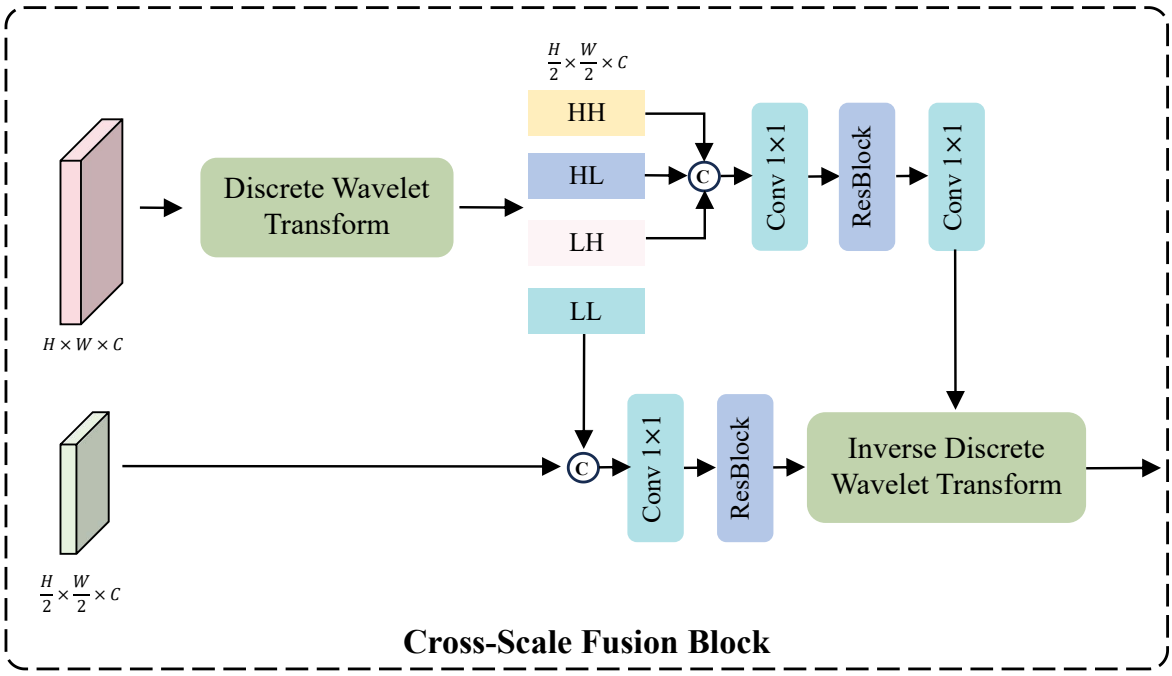


Figure 5. The architecture of the proposed CSFM. CSFM is a cross-scale fusion module that leverages wavelet transforms to effectively merge feature maps of different spatial resolutions.

3.6. Training Loss

Due to the typical display of HDR images after tonemapping, training the network on tonemapped images is more effective than training directly in the HDR domain. When provided with an HDR image H in the HDR domain, we compress the image's range using the μ-law transformation.

$$T(H) = \frac{\log(1+\mu H)}{\log(1+\mu)}, \tag{21}$$

where μ represents a parameter that defines the degree of compression, and $T(H)$ represents the tonemapped image. Throughout this work, we maintain H within the range [0, 1] and set μ to 5000.

$\hat{H}$ is the predicted result obtained from our HL-HDR model, and H is the Ground Truth. Here, we employ L_1 loss to compute the loss. Additionally, we use an auxiliary perceptual loss L_p for supervision [14]. The perceptual loss measures the difference between the output image and the Ground Truth image in the feature representations of multiple layers in a pre-trained CNN, achieved by computing the mean squared error between the feature maps of each layer. We can express this as follows:

$$L_1 = \|T(H) - T(\hat{H})\|_1, \tag{22}$$

$$L_p = \|\theta_j(T(H)) - \theta_j(T(\hat{H}))\|_1, \tag{23}$$

where θ_j represents the j^{th} convolutional feature extracted from the pre-trained VGG-16 network, with j denoting the j^{th} layer.

Therefore, our final loss function is the result of adding L_1 and L_p, with different weights assigned to each, for which we introduce a parameter λ. The final loss function can be expressed by the following formula:

$$L_{total} = L_1 + \lambda L_p, \tag{24}$$

where λ is set to 0.01.

 doi:10.20944/preprints202510.1208.v1

4. Experiments

4.1. Experiments Settings

Datasets. The proposed method has been trained on three distinct datasets: Kalantari's dataset [8], Tel's dataset [15], and Hu's dataset [31]. Kalantari's dataset consists of 74 training samples and 15 testing samples captured from real-world scenes, with exposure values set at {-2, 0, +2} and {-3, 0, +3}. Tel's dataset comprises 108 training samples and 36 testing samples. For Hu's dataset, the first 85 samples were used for training, while the remaining 15 were reserved for testing. This dataset employs an exposure bias of {-2, 0, +2} and is synthetically generated using a game engine sensor. To evaluate the effectiveness and generalization capability of the proposed model, we conducted tests on Sen's dataset [7] and Tursun's dataset [32] using weights pre-trained on Kalantari's dataset. Since these two datasets contain only LDR images at different exposure levels and lack ground truth, the performance comparison across methods is limited to subjective assessment.

Evaluation Metrics. We use five objective measures for quantitative comparison: PSNR-μ, SSIM-μ, PSNR-l, SSIM-l, and HDR-VDP-2 [33]. Here, μ and L indicate that the metrics are calculated in the tonemapped domain and the linear domain, respectively.

Implementation Details. Our implementation is based on PyTorch. Before training, we sample 256×256 patches from the dataset with a stride of 64. To enhance the diversity of the training data, we apply data augmentation techniques including rotation and flipping, as well as their combinations, resulting in six different augmentation strategies. We employ the Adam optimizer with a batch size of 8 and an initial learning rate of 2×10^{-4}, which is reduced every 70 epochs. The model is trained for a total of 250 epochs on a single NVIDIA GeForce RTX 4090 GPU.

4.2. Comparison with the State-of-the-art Methods

To comprehensively evaluate the performance of our model, we compared it against representative state-of-the-art deep learning-based approaches spanning different architectural paradigms. Specifically, we considered six CNN-based methods, including DHDR [9], AHDR [10], NHDRR [20], APNT [34], PGN [35], and SAFNet [26]; one GAN-based method, HDR-GAN [36]; three Transformer-based models, namely CA-ViT [14], SCTNet [15], and HyHDR [22]; as well as two diffusion-based methods, DiffHDR [37] and LFDiff [13].

Datasets w/ Ground Truth. Table 1, 2, and 3 presents the quantitative results of HL-HDR on three datasets. Our method is compared against several state-of-the-art approaches using the testing data from [8], [31], and [15], which consist of challenging samples characterized by saturated backgrounds and foreground motions. The average of all quantitative results is computed across the testing images. Notably, our method performs remarkably well on Kalantari's dataset [8], achieving state-of-the-art performance in PSNR-μ, along with competitive results in other metrics. On Hu's dataset [31], our method achieves strong performance in both PSNR-μ and PSNR-l, with PSNR-l outperforming the second-best approach by 0.82 dB. On Tel's dataset [15], our method demonstrates overall superiority, where both PSNR-μ and PSNR-l substantially surpass the second-best approach, with gains of 0.62 dB and 0.32 dB, respectively.

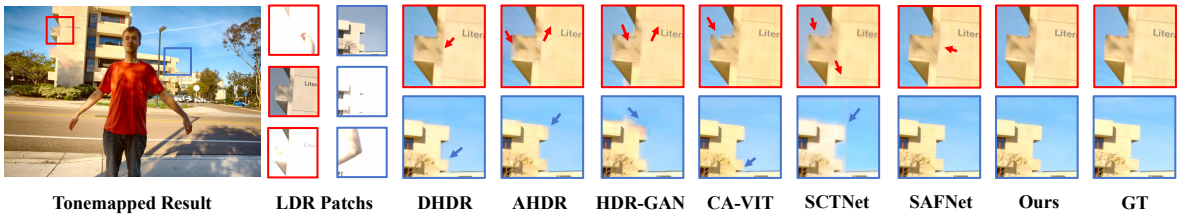
In Figure 6 (a)(b)(c), the datasets present significant challenges due to large-scale foreground motion and severe over/under-exposed regions. We qualitatively compare our method with several state-of-the-art approaches. Most competing methods suffer from ghosting artifacts in regions with motion and saturation. On Kalantari's dataset[8], DHDR [9] exhibits severe ghosting, while AHDR [10], HDR-GAN [36], and SCTNet [15] not only fail to recover complete structural information but also perform poorly in detail restoration. For example, in the patch comparison shown in Figure 6 (a), these three methods fail to reconstruct the balcony, with sky elements incorrectly blended in, and the wall textures appear blurry. CA-VIT [14] and SCTNet [15] further suffer from blocky ghosting artifacts due to patch-based sampling. In contrast, SAFNet [26] exhibits only slight wall deformation. Our proposed method not only restores the overall structural content accurately but also excels in preserving fine

 doi:10.20944/preprints202510.1208.v1

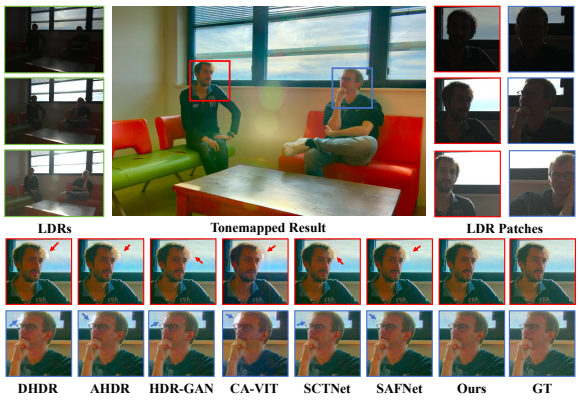
Table 1. Quantitative comparisons on Kalantari's dataset [8]. The best results are highlighted in red.

Methods	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l	HDR-VDP-2
DHDR [9]	41.64	40.91	0.9869	0.9858	60.50
AHDR [10]	43.62	41.03	0.9900	0.9862	62.30
NHDRR [20]	42.41	41.08	0.9887	0.9861	61.21
HDR-GAN [36]	43.92	41.57	0.9905	0.9865	65.45
APNT [34]	43.94	41.61	0.9898	0.9879	64.05
CA-ViT [14]	44.32	42.18	0.9916	0.9884	66.03
HyHDR [22]	44.64	42.47	0.9915	0.9894	66.05
DiffHDR [37]	44.11	41.73	0.9911	0.9885	65.52
SCTNet [15]	44.43	42.21	0.9918	0.9891	66.64
PGN [35]	44.73	42.27	0.9918	0.9890	66.08
SAFNet [26]	44.66	43.18	0.9919	0.9901	66.11
LFDiff [13]	44.76	42.59	0.9919	0.9906	66.54
Ours	44.81	42.69	0.9921	0.9901	66.71

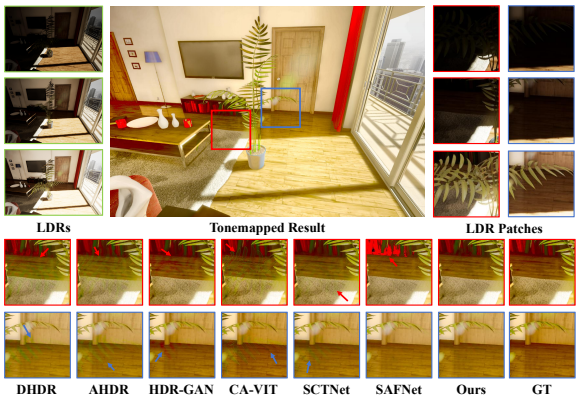
details. In particular, the wall lines remain sharp and clear, demonstrating the strong capability of our model in capturing and restoring fine-grained information.



(a) Examples of Kalantari *et al.*'s dataset [8]



(b) Examples of Tel *et al.*'s dataset [15]



(c) Examples of Hu *et al.*'s dataset [31]

Figure 6. Examples of three representative HDR datasets: (a) Kalantari *et al.*'s dataset, (b) Tel *et al.*'s dataset, and (c) Hu *et al.*'s dataset.

In Figure 6 (b), we show a comparison scene from Tel's dataset [15], where only the heads of two people exhibit slight motion. All other methods, however, produced noticeable ghosting artifacts in these motion regions. In contrast, our method accurately detects the motion areas and achieves superior image reconstruction. In Figure 6 (c), we show a comparison scene from Hu's dataset [31],

 doi:10.20944/preprints202510.1208.v1

Table 2. Quantitative comparisons on Tel's dataset [15]. The best results are highlighted in red.

Methods	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l	HDR-VDP-2
DHDR [9]	40.05	43.37	0.9794	0.9924	67.09
AHDR [10]	42.08	45.30	0.9837	0.9943	68.80
NHDRR [20]	36.68	39.61	0.9590	0.9853	65.41
HDR-GAN [36]	41.71	44.87	0.9832	0.9949	69.57
CA-ViT [14]	42.39	46.35	0.9844	0.9948	69.23
SCTNet [15]	42.55	47.51	0.9850	0.9952	70.66
DiffHDR [37]	42.18	45.63	0.9841	0.9946	69.88
SAFNet [26]	42.68	47.46	0.9792	0.9955	68.16
Ours	43.30	47.83	0.9878	0.9957	70.73

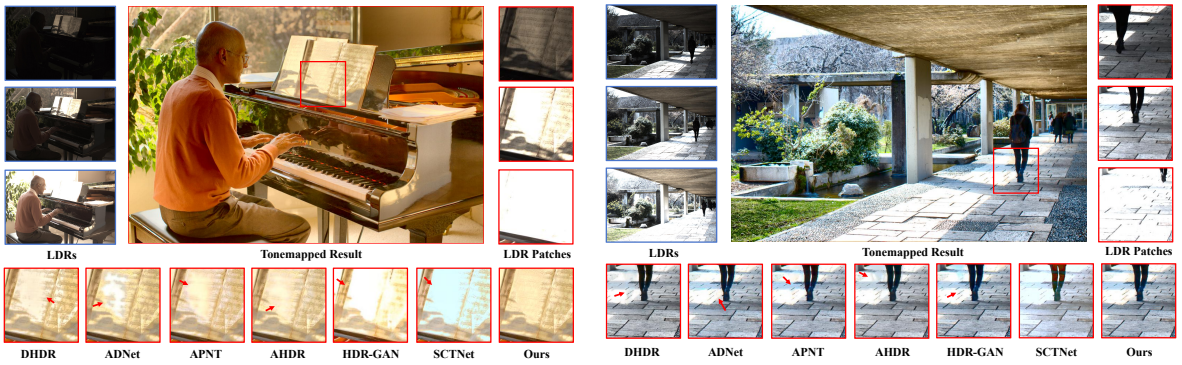
Table 3. Quantitative comparisons on Hu's dataset [31]. The best results are highlighted in red.

Methods	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l	HDR-VDP-2
DHDR [9]	41.13	41.20	0.9870	0.9941	70.82
AHDR [10]	45.76	49.22	0.9956	0.9980	75.04
NHDRR [20]	45.15	48.75	0.9956	0.9981	74.86
HDR-GAN [36]	45.86	49.14	0.9945	0.9989	75.19
APNT [34]	46.41	47.97	0.9953	0.9986	73.06
CA-ViT [14]	48.10	51.17	0.9947	0.9989	77.12
HyHDR [22]	48.46	51.91	0.9959	0.9991	77.24
DiffHDR [37]	48.03	50.23	0.9954	0.9989	76.22
SCTNet [15]	48.10	51.03	0.9963	0.9991	77.14
PGN [35]	48.66	52.49	0.9965	0.9992	77.33
SAFNet [26]	47.18	49.35	0.9951	0.9990	76.83
LFDiff [13]	48.74	52.10	0.9968	0.9993	77.35
Ours	49.02	52.92	0.9970	0.9992	77.55

where the motion is much more substantial. Except for our method, all other approaches generated large ghosting regions in the motion areas, significantly degrading the visual quality.

Evaluation on Datasets without Ground Truth. To evaluate the generalization capability of the proposed HDR imaging method, we tested the model trained on Kalantari's Dataset [8] on Sen's Dataset [7] and Tursun's Dataset [32], both of which lack ground truth. Consequently, the quality of the generated HDR images can only be assessed through subjective visual inspection. Notably, in Figure 7 (a), most methods fail to recover overexposed regions, whereas our method performs exceptionally well, not only avoiding overexposure but also successfully restoring rich detail. In Figure 7 (b), both our method and SCTNet are visually the most appealing, with no noticeable ghosting caused by human motion. This is because the scene contains abundant background information, and Transformer-based methods can fully exploit long-range dependencies to extract information from similar regions, thereby restoring details in motion areas and generating ghost-free images.

 doi:10.20944/preprints202510.1208.v1



(a) Example from Sen *et al.*'s dataset (b) Example from Tursen *et al.*'s dataset

Figure 7. Examples of Sen *et al.*'s dataset [7] and Tursen *et al.*'s dataset [32].

4.3. Ablation Studies

We conducted ablation experiments on the Kalantari dataset to evaluate the effectiveness of each module in our model. The following sections present the ablation analysis from three perspectives, corresponding to the main components of the model.

4.3.1. Effect of Different Alignment Modules

To validate the effectiveness of the proposed Frequency Alignment Module, we compared it with two alternative models: one without any alignment module, and the other using the AHDR [10] alignment module. All other model components and parameters were kept identical.

As shown in Table 4, it is evident that the model without any alignment module achieves the lowest metrics. The model using AHDR [10] for alignment shows a slight improvement, but the gain is limited. In contrast, when aligned using our proposed FAM, both PSNR-μ and PSNR-l increase significantly, with PSNR-l rising by 0.63dB. Although the model without alignment achieves relatively lower scores, it still outperforms several existing methods such as CA-ViT [14] and SCTNet [15], indirectly highlighting the effectiveness of our proposed FDPB. Furthermore, we provide a visual comparison. In Figure 8, we select an example from the test set that contains substantial motion and overexposed regions. The model without alignment exhibits large ghosting artifacts, the AHDR-aligned model shows some improvement, while our FAM-aligned result almost entirely eliminates ghosting. This is attributed to our use of optical flow to capture large-scale motion, which further enhances image reconstruction.

Table 4. Comparison of different alignment modules. The best results are highlighted in red.

Alignment Module	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l
None	44.53	42.06	0.9917	0.9890
AHDR	44.62	42.22	0.9919	0.9892
FAM (Ours)	44.81	42.69	0.9921	0.9906

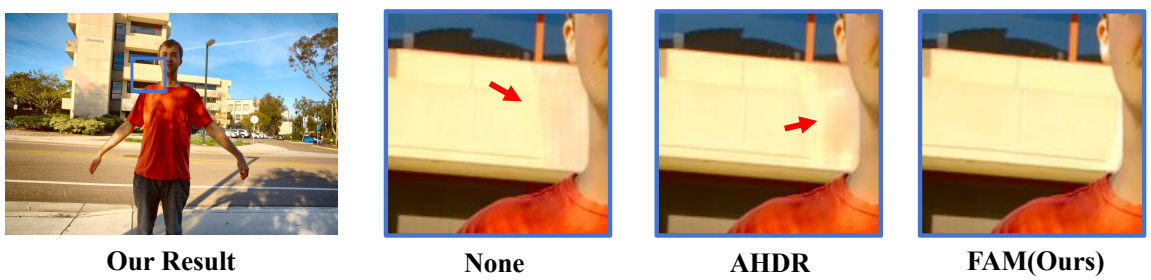


Figure 8. Comparison of different alignment strategies.

 doi:10.20944/preprints202510.1208.v1

4.3.2. Ablation Analysis of Components in the Frequency Alignment Module

To validate the rationale behind our proposed FAM, which decomposes high and low-frequency features and processes them using different methods, we designed four experimental settings: (1) No high-low frequency separation, aligning the two frames using only optical flow; (2) No high-low frequency separation, aligning the two frames using convolution and attention(AHDR-aligned); (3) High-low frequency separation, aligning both high and low-frequency features using optical flow; (4) High-low frequency separation, aligning both high and low-frequency features using convolution and attention.

Table 5 presents the comparison results of different alignment strategies for high and low-frequency features. As shown in the table, the non-separation strategies (1) and (2) exhibit notable differences: the optical-flow-based approach achieves significantly better PSNR-l and SSIM-l compared to the convolution + attention approach, demonstrating the superiority of optical flow in handling large-scale motion. When high and low-frequency features are separated and both aligned using optical flow (3), the performance is comparable to (1) but with only marginal improvement, indicating that separation alone does not yield substantial benefits. In contrast, fully relying on convolution and attention after separation (4) performs even worse than the non-separated cases, highlighting its limitations in capturing large-scale motion. In comparison, our proposed FAM achieves the best results in PSNR-μ, SSIM-μ, and SSIM-l, while maintaining overall stable performance. These results validate the effectiveness of combining optical flow with frequency separation in our design.

Table 5. Comparison of different alignment strategies for high and low-frequency features. The best results are highlighted in red.

Method	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l
(1) No separation, optical flow	44.67	42.81	0.9920	0.9904
(2) No separation, conv + attention	44.62	42.22	0.9919	0.9892
(3) Separation, optical flow	44.65	42.67	0.9920	0.9896
(4) Separation, conv + attention	44.51	42.14	0.9920	0.9898
FAM (Ours)	44.81	42.69	0.9921	0.9906

4.3.3. Ablation Analysis of Components in the Frequency Decomposition Processing Block

To validate the effectiveness of the proposed FDPB, we design a more fine-grained ablation study consisting of six comparative schemes: (1) Without frequency decomposition, the aligned features are simultaneously fed into both the Global Feature Extractor (GFE) and the Local Feature Extractor (LFE); (2) With frequency decomposition, but with swapped branch functions, where the high-frequency features are processed by the Global Feature Extractor (GFE) and the low-frequency features are processed by the Local Feature Extractor (LFE); (3) With frequency decomposition, but extracting both high and low-frequency features using the LFE; (4) With frequency decomposition, but extracting both high and low-frequency features using the GFE; (5) Removing the dense residual connections in the low-frequency branch; (6) Removing the wavelet-based Cross-Scale Fusion Module at each layer of the high-frequency branch.

Table 6 presents a comprehensive ablation study of the proposed Frequency Decomposition Processing Block (FDPB). It can be observed that performing frequency decomposition significantly enhances HDR reconstruction performance: comparing the scheme without decomposition (1) to the schemes with decomposition (3) and (4), PSNR-l is generally improved, indicating that separating high and low-frequency features facilitates more effective extraction of global structures and local details. Notably, although scheme (4) achieves the highest PSNR-l, its PSNR-μ is relatively low due to insufficient processing of high-frequency information, demonstrating that relying solely on the Global Feature Extractor (GFE) is inadequate for restoring image details. The importance of matching feature types to the appropriate extractor is highlighted in scheme (2), where swapping the high-frequency and low-frequency branches leads to a noticeable performance drop, showing that high-frequency

 doi:10.20944/preprints202510.1208.v1

Table 6. Ablation study of the proposed FDPB. The best results are highlighted in red.

Method	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l
(1) No decomposition, GFE+LFE	44.49	42.61	0.9919	0.9896
(2) Frequency decomposition, swapped	44.45	42.33	0.9918	0.9895
(3) Frequency decomposition, LFE only	44.36	42.22	0.9918	0.9892
(4) Frequency decomposition, GFE only	44.64	42.88	0.9920	0.9900
(5) Low-freq w/o dense res.	44.68	42.59	0.9920	0.9902
(6) High-freq w/o CSFM	44.61	42.56	0.9919	0.9901
FDPB (Ours)	44.81	42.69	0.9921	0.9906

features are better handled by the Local Feature Extractor (LFE) and low-frequency features by the GFE. Furthermore, removing dense residual connections in the low-frequency branch (5) or the Cross-Scale Fusion Module in the high-frequency branch (6) results in decreased performance, emphasizing the critical role of these components in preserving structural information and enhancing details. Overall, the complete FDPB design, integrating frequency decomposition, proper branch assignment, dense residual connections, and wavelet-based cross-scale fusion, achieves the best results across all metrics, confirming its effectiveness in restoring high-quality, ghost-free HDR images.

5. Conclusions

In this paper, we presented HL-HDR, a novel high-low frequency-aware HDR reconstruction network tailored for dynamic scenes. By explicitly decomposing and processing features into high- and low-frequency components, HL-HDR effectively integrates global context modeling with fine-grained detail restoration. The proposed Frequency Alignment Module (FAM) enables accurate motion handling and structure preservation, while the Frequency Decomposition Processing Block (FDPB) achieves hierarchical feature fusion across scales. Extensive experiments on multiple public HDR benchmarks validate that HL-HDR delivers superior ghost-free HDR results compared to existing approaches. In future work, we plan to further explore lightweight designs and real-time deployment to extend the applicability of HL-HDR in practical scenarios.

Acknowledgments: This work is supported by Science and Technology Development Program Projects of the Housing and Urban-Rural Development Department of Shaanxi Province (No.2023-k48).

References

1. Bogoni, L. Extending dynamic range of monochrome and color images through fusion. In Proceedings of the Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. IEEE, 2000, Vol. 3, pp. 7–12.
2. Tomaszewska, A.; Mantiuk, R. Image registration for multi-exposure high dynamic range image acquisition **2007**.
3. Grosch, T.; et al. Fast and robust high dynamic range image generation with camera and object movement. *Vision, Modeling and Visualization, RWTH Aachen* **2006**, *277284*, 2.
4. Lee, C.; Li, Y.; Monga, V. Ghost-free high dynamic range imaging via rank minimization. *IEEE signal processing letters* **2014**, *21*, 1045–1049.
5. Pece, F.; Kautz, J. Bitmap movement detection: HDR for dynamic scenes. In Proceedings of the 2010 Conference on Visual Media Production. IEEE, 2010, pp. 1–8.
6. Hu, J.; Gallo, O.; Pulli, K.; Sun, X. HDR deghosting: How to deal with saturation? In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 1163–1170.
7. Sen, P.; Kalantari, N.K.; Yaesoubi, M.; Darabi, S.; Goldman, D.B.; Shechtman, E. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.* **2012**, *31*, 203–1.
8. Kalantari, N.K.; Ramamoorthi, R.; et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.* **2017**, *36*, 144–1.

 doi:10.20944/preprints202510.1208.v1

9. Wu, S.; Xu, J.; Tai, Y.W.; Tang, C.K. Deep high dynamic range imaging with large foreground motions. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 117–132.
10. Yan, Q.; Gong, D.; Shi, Q.; Hengel, A.v.d.; Shen, C.; Reid, I.; Zhang, Y. Attention-guided network for ghost-free high dynamic range imaging. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1751–1760.
11. Zhang, X.; Hu, T.; He, J.; Yan, Q. Efficient content reconstruction for high dynamic range imaging. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024, pp. 7660–7664.
12. Zhang, X.; Zhu, Q.; Hu, T.; Yan, Q. EiffHDR: An Efficient Network for Multi-Exposure High Dynamic Range Imaging. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024, pp. 6560–6564.
13. Hu, T.; Yan, Q.; Qi, Y.; Zhang, Y. Generating Content for HDR Deghosting from Frequency View. *arXiv preprint arXiv:2404.00849* **2024**.
14. Liu, Z.; Wang, Y.; Zeng, B.; Liu, S. Ghost-free high dynamic range imaging with context-aware transformer. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 344–360.
15. Tel, S.; Wu, Z.; Zhang, Y.; Heyrman, B.; Demonceaux, C.; Timofte, R.; Ginhac, D. Alignment-free HDR Deghosting with Semantics Consistent Transformer. *arXiv preprint arXiv:2305.18135* **2023**.
16. Zhang, X.; Chen, G.; Hu, T.; Yang, K.; Zhang, F.; Yan, Q. HL-HDR: Multi-Exposure High Dynamic Range Reconstruction with High-Low Frequency Decomposition. In Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN). IEEE, 2024, pp. 1–9.
17. Kang, S.B.; Uyttendaele, M.; Winder, S.; Szeliski, R. High dynamic range video. *ACM Transactions on Graphics (TOG)* **2003**, *22*, 319–325.
18. Jacobs, K.; Loscos, C.; Ward, G. Automatic high-dynamic range image generation for dynamic scenes. *IEEE Computer Graphics and Applications* **2008**, *28*, 84–93.
19. Ma, K.; Li, H.; Yong, H.; Wang, Z.; Meng, D.; Zhang, L. Robust multi-exposure image fusion: a structural patch decomposition approach. *IEEE Transactions on Image Processing* **2017**, *26*, 2519–2532.
20. Yan, Q.; Zhang, L.; Liu, Y.; Zhu, Y.; Sun, J.; Shi, Q.; Zhang, Y. Deep HDR imaging via a non-local network. *IEEE Transactions on Image Processing* **2020**, *29*, 4308–4322.
21. Song, J.W.; Park, Y.I.; Kong, K.; Kwak, J.; Kang, S.J. Selective TransHDR: Transformer-Based Selective HDR Imaging Using Ghost Region Mask. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 288–304.
22. Yan, Q.; Chen, W.; Zhang, S.; Zhu, Y.; Sun, J.; Zhang, Y. A Unified HDR Imaging Method with Pixel and Patch Level. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22211–22220.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
24. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5728–5739.
25. Yan, Q.; Zhang, S.; Chen, W.; Liu, Y.; Zhang, Z.; Zhang, Y.; Shi, J.Q.; Gong, D. A lightweight network for high dynamic range imaging. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 824–832.
26. Kong, L.; Li, B.; Xiong, Y.; Zhang, H.; Gu, H.; Chen, J. Safnet: Selective alignment fusion network for efficient hdr imaging. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 256–273.
27. Dong, J.; Pan, J.; Yang, Z.; Tang, J. Multi-Scale Residual Low-Pass Filter Network for Image Deblurring. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 12345–12354.
28. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646–1654.
29. Pan, J.; Liu, S.; Sun, D.; Zhang, J.; Liu, Y.; Ren, J.; Li, Z.; Tang, J.; Lu, H.; Tai, Y.W.; et al. Learning dual convolutional neural networks for low-level vision. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3070–3079.

 doi:10.20944/preprints202510.1208.v1

30. Huang, J.J.; Dragotti, P.L. WINNet: Wavelet-inspired invertible network for image denoising. *IEEE Transactions on Image Processing* **2022**, *31*, 4377–4392.
31. Hu, J.; Choe, G.; Nadir, Z.; Nabil, O.; Lee, S.J.; Sheikh, H.; Yoo, Y.; Polley, M. Sensor-realistic synthetic data engine for multi-frame high dynamic range photography. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 516–517.
32. Tursun, O.T.; Akyüz, A.O.; Erdem, A.; Erdem, E. An objective deghosting quality metric for HDR images. In Proceedings of the Computer Graphics Forum. Wiley Online Library, 2016, Vol. 35, pp. 139–152.
33. Mantiuk, R.; Kim, K.J.; Rempel, A.G.; Heidrich, W. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)* **2011**, *30*, 1–14.
34. Chen, J.; Yang, Z.; Chan, T.N.; Li, H.; Hou, J.; Chau, L.P. Attention-guided progressive neural texture fusion for high dynamic range image restoration. *IEEE Transactions on Image Processing* **2022**, *31*, 2661–2672.
35. Yan, Q.; Yang, K.; Hu, T.; Chen, G.; Dai, K.; Wu, P.; Ren, W.; Zhang, Y. From dynamic to static: Stepwisely generate HDR image for ghost removal. *IEEE Transactions on Circuits and Systems for Video Technology* **2024**.
36. Niu, Y.; Wu, J.; Liu, W.; Guo, W.; Lau, R.W. HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. *IEEE Transactions on Image Processing* **2021**, *30*, 3885–3896.
37. Yan, Q.; Hu, T.; Sun, Y.; Tang, H.; Zhu, Y.; Dong, W.; Van Gool, L.; Zhang, Y. Toward high-quality HDR deghosting with conditional diffusion models. *IEEE Transactions on Circuits and Systems for Video Technology* **2023**, *34*, 4011–4026.