electronics

MDPI

Article

Attention-Guided HDR Reconstruction for Enhancing Smart City Applications

Yung-Yao Chen [1], Chih-Hsien Hsia [2,*], Sin-Ye Jhong [3,*] and Chin-Feng Lai [3]

[1] Department of Electronics and Computer Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan; yungyaochen@gapps.ntust.edu.tw
[2] Department of Computer Science and Information Engineering, National Ilan University, Yilan 260, Taiwan
[3] Department of Engineering Science, National Cheng Kung University, Tainan 701, Taiwan; cinfon@ncku.edu.tw
* Correspondence: hsiach@niu.edu.tw (C.-H.H.); n98081034@ncku.edu.tw (S.-Y.J.)

Abstract: In the context of smart city development, video surveillance serves as a critical component for maintaining public safety and operational efficiency. However, traditional surveillance systems are often constrained by a limited dynamic range, leading to the loss of essential image details. To address this limitation, this paper introduces HDRFormer, an innovative framework designed to enhance high dynamic range (HDR) image quality in edge–cloud-based video surveillance systems. Leveraging advanced deep learning algorithms and Internet of Things (IoT) technology, HDRFormer employs a unique architecture comprising a feature extraction module (FEM) and a weighted attention module (WAM). The FEM leverages a transformer-based hierarchical structure to adeptly capture multi-scale image information. In addition, the guided filters are utilized to steer the network, thereby enhancing the structural integrity of the images. On the other hand, the WAM focuses on reconstructing saturated areas, improving the perceptual quality of the images, and rendering the reconstructed HDR images with naturalness and color saturation. Extensive experiments on multiple HDR image reconstruction datasets demonstrate HDRFormer's substantial improvements, achieving up to a 2.7 dB increase in the peak signal-to-noise ratio (PSNR) and an enhancement of 0.09 in the structural similarity (SSIM) compared to existing methods. In addition, the framework exhibits outstanding performance in multi-scale structural similarity (MS-SSIM) and HDR visual difference predictor (HDR-VDP2.2). The proposed method not only outperforms the existing HDR reconstruction techniques but also offers better generalization capabilities, laying a robust foundation for future applications in smart cities.

Keywords: smart city; surveillance system; high dynamic range; vision transformer; attention mechanism; guided filter; image reconstruction

check for updates

Citation: Chen, Y.-Y.; Hsia, C.-H.; Jhong, S.-Y.; Lai, C.-F. Attention-Guided HDR Reconstruction for Enhancing Smart City Applications. *Electronics* **2023**, *12*, 4625. https://doi.org/10.3390/electronics12224625

Academic Editors: Dah-Jye Lee, Wenfeng Zheng, Mingzhe Liu, Chao Liu and Dan Wang

Received: 13 October 2023
Revised: 7 November 2023
Accepted: 10 November 2023
Published: 12 November 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In our technologically advanced society, innovations are rapidly reshaping urban environments, transitioning from foundational elements such as smart homes and factories into comprehensive smart cities. These cities, equipped with advanced surveillance, healthcare, and intelligent transportation systems, enhance daily life quality [1]. By leveraging soft computing, deep learning, and computer vision, artificial intelligence (AI) efficiently processes the vast amounts of data generated by IoT devices, bringing the once abstract idea of a "smart city" to life [2]. A pivotal component of this intelligent ecosystem is video surveillance [3]. Traditional surveillance systems, while proficient in autonomous searching, detection, and tracking, grapple with challenges arising from atmospheric visibility and sensor limitations [4]. In the real world, images span a broad luminance spectrum, often surpassing human visual capabilities. Many camera systems, bound by their limited luminance range, produce subpar image quality [5]. Standard cameras with low dynamic

range (LDR) sensors frequently falter under fluctuating lighting scenarios, resulting in the loss of crucial image details in underexposed and overexposed regions, undermining the system's detection and recognition abilities [6].

Edge computing has arisen as a powerful alternative to cloud computing. By processing data directly at its source, edge computing mitigates latency and congestion issues [7,8]. When integrated with cloud computing, a robust edge–cloud computing framework emerges. Within this framework, data from scattered edge devices are amalgamated in the cloud, which then provides global insights back to these devices, enhancing real-time video surveillance capabilities [9]. The integration of high dynamic range (HDR) technology can further enhance this system, ensuring surveillance tasks benefit from superior image quality, even in challenging lighting conditions. The architecture of the edge–cloud-based HDR surveillance system is depicted in Figure 1. This collaborative approach between edge and cloud computing, bolstered by HDR technology, is poised to redefine next-generation video surveillance, emphasizing the importance of data security and privacy protection in smart homes, industries, and cities.

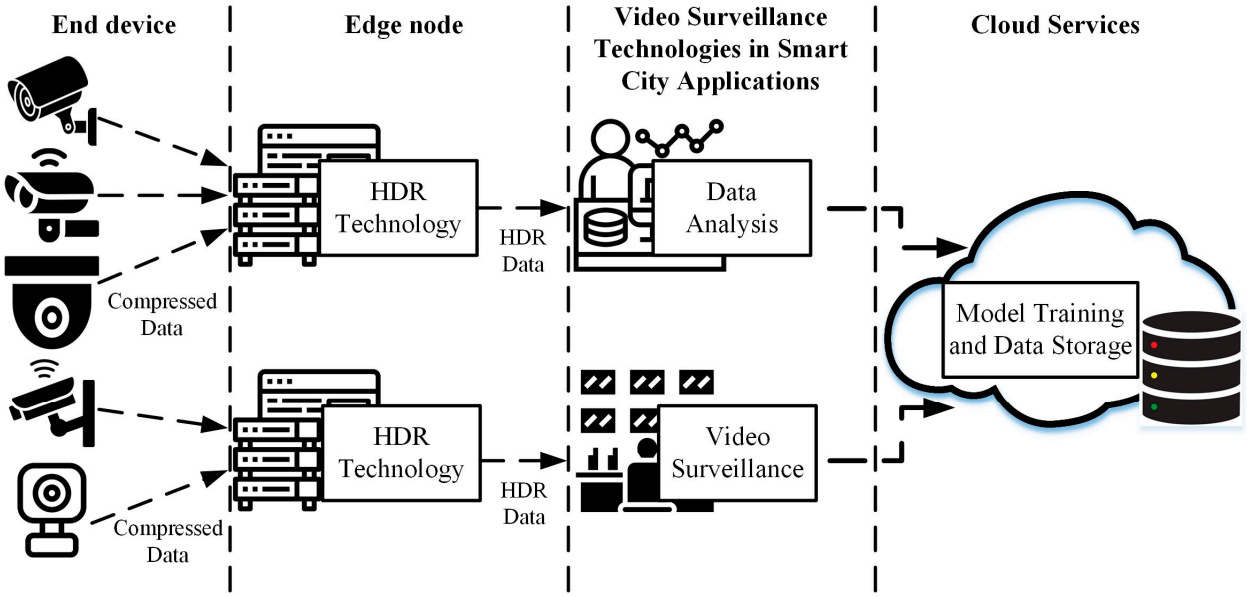


Figure 1. The architecture of the edge–cloud-based HDR surveillance system. By deploying HDR for a priori processing at the edge node, the reliability of subsequent video surveillance technology is enhanced, thereby providing a robust foundation for applications in smart cities.

In the realm of imaging, high dynamic range (HDR) stands out for its ability to vividly capture the richness of real-world scenes. However, a significant challenge arises when using consumer-grade camera sensors. These sensors often grapple with capturing the luminance and contrast of real-world scenes, leading to information loss in either overexposed or underexposed areas. To address this, researchers have explored hardware and software solutions, leading to the development of HDR imaging techniques. There are two primary methods to achieve HDR images. The first involves direct capture using specialized hardware. The second method, on the other hand, leans towards artificial reconstruction from LDR images, which are typically captured with standard cameras. This reconstruction can be achieved through both handcrafted and data-driven techniques. It is worth noting that while HDR-specific cameras have seen significant advancements, their high costs often act as a barrier to mass adoption.

Given these constraints, the focus has shifted towards HDR image reconstruction as a more viable alternative. The most common method involves capturing a series of LDR images at different exposures and then merging them using multi-exposure fusion techniques [10,11]. In this method, one image, typically the one with median exposure, is

used as a reference. The other images serve to fill in the details for areas that are either overexposed or underexposed. The crux of this method lies in the perfect alignment of LDR images at the pixel level, which, when achieved, results in high-quality HDR images. However, challenges arise when there is object movement, causing misalignment and leading to artifacts like blurring and ghosting in the resultant HDR image. An alternative to this is the single-image HDR reconstruction technique [12,13]. This method aims to expand the dynamic range derived from just one LDR image, eliminating issues such as ghosting and making it suitable for static and dynamic scenes. However, this task is challenging. Using only a single LDR image, it is difficult to compensate for the extensive exposure information that multi-exposure methods offer. In addition, LDR images, especially in saturated regions, often undergo significant content loss, making the reconstruction of authentic real-world information a daunting task.

Traditional single-image HDR reconstruction methods [14,15] primarily enhance brightness or contrast directly through linear or nonlinear functions or employ local analysis methods combined with expansion maps. However, a significant limitation of these handcrafted methods is their reliance on fixed parameters, which are often determined by environmental variables, thereby restricting their adaptability. Recently, deep learning-based single-image HDR reconstruction methods have gained popularity, producing convincing HDR reconstruction results [12,13]. These methods typically utilize one or multiple encoder–decoder architectures to model the transformation from LDR to HDR images. However, they face several challenges. For instance, while these models are adept at capturing multi-scale information, up-sampling and down-sampling often lead to a loss of spatial information. Furthermore, convolution operations in CNNs inherently face difficulties in modeling long-range dependencies, which can result in outputs exhibiting unnatural tones and contrasts. Recently, the transformer [16] model for natural language processing (NLP) has been adapted to the computer vision domain. The transformer's multi-head self-attention (MSA) module excels at learning non-local similarities and long-range dependencies, presenting a potential solution to the shortcomings of CNN-based methods [17,18]. However, it is worth noting that while deep learning offers multiple solutions, most HDR reconstructions rely on mean squared error (MSE) solutions. The inherent nature of MSE, which tends to average out many solutions, renders it less reliable, especially when restoring details in areas with sparse saturation.

To address these challenges, we present the HDRFormer, an IoT-based HDR framework. The HDRFormer comprises two main components: the feature extraction module (FEM) and the weighted attention module (WAM). The FEM, designed with a hierarchical structure, is adept at extracting multi-scale information. Recognizing the importance of the structural information of the reconstructed image, we adopt an edge-guided strategy to reinforce edge structures. By preserving image edge information through guided filters and enhancing structural performance through convolution operations, we found that directly replacing convolution layers with transformer components is not optimal. Although transformers can capture global dependencies, edge pixels cannot utilize adjacent pixels outside the patch for restoration during computation, potentially leading to artifacts and compromising reconstruction quality. Therefore, we introduce the LASwin transformer module, which employs self-attention based on non-overlapping windows, significantly reducing computational demands. This module also incorporates convolution operations to capture the local context of images better, thereby enhancing HDR reconstruction capabilities. Furthermore, research [19] indicates that the perceptual quality of images deteriorates significantly when details in saturation areas are lacking. Recognizing the importance of these areas for visual quality, we designed the WAM to accentuate information in saturation areas. The WAM utilizes a supervised mechanism with a weight map, prioritizing the reconstruction of saturation areas. This is further complemented by an attention map, which helps refine details in these critical regions.

The main contributions of this study are as follows:

(1) A novel HDR reconstruction framework designed specifically for enhancing image quality in edge–cloud-based video surveillance systems. This model integrates a unique FEM and a WAM to produce HDR images of exceptional quality across diverse scenarios.
(2) Within the FEM, this study introduces a hybrid architecture that melds the strengths of self-attention mechanisms with convolution operations. This approach adeptly captures global image dependencies while retaining essential local image contexts, ensuring optimal HDR image reconstruction.
(3) Acknowledging the significance of saturated areas in HDR images, the WAM is equipped with a supervised weight map and attention map. This design emphasizes the reconstruction of these pivotal areas, refining intricate details and boosting the perceptual quality of the resultant images, even under challenging lighting conditions.
(4) We conducted an extensive series of experiments on multiple publicly available HDR image reconstruction datasets. The quantitative and qualitative results emphatically underscore the superior performance and robustness of the proposed method.

The remainder of this paper is organized as follows: Section 2 reviews related works. Section 3 details the proposed model structure and loss functions. Section 4 describes the datasets used and presents the qualitative and quantitative comparisons between the proposed and existing methods. Finally, Section 5 concludes our work and discusses potential avenues for future research.

2. Related Studies

2.1. Single-Image HDR Reconstruction

Single-image HDR reconstruction, commonly referred to as inverse tone mapping (ITM) [14], has undergone extensive research over the past decades. While multi-image HDR reconstruction techniques might introduce ghosting and blurring artifacts, deriving details from a single-exposure LDR image in single-image HDR remains a formidable challenge. Early ITM approaches can be categorized into global and local techniques. Global methods, such as those proposed by Landis et al. [15] and Bist et al. [20], employ power and gamma functions for content expansion, respectively. However, these methods demand accurate parameter configurations. On the other hand, local techniques, as demonstrated by Banterle et al. [14] and Didyk et al. [21], employ analytical methods combined with expansion maps to enhance the luminance range of highlighted regions. Rempel et al. [22] apply Gaussian filtering and image pyramids to retrieve information from saturated regions. Traditional ITM techniques face the challenge of identifying a universal function applicable to all images, prompting the introduction of deep CNN-based solutions. Eilertsen et al. [13] and Wu et al. [23] adopt the U-Net and MobileNetV2 architectures, respectively, to estimate the content in overexposed regions and integrate it with underexposed regions to produce HDR images. Liu et al. [12] present a novel approach that learns to reverse the camera pipeline to generate HDR images. Marnerides et al. [24] advocate for a multi-branch CNN architecture, while Khan et al. [25] implement a recursive neural network. However, these methods sometimes fall short in restoring pixels in saturated regions. In contrast, the proposed method leverages a WAM to precisely enhance these saturated regions.

2.2. Attention Mechanism

Attention mechanisms have emerged as a crucial component in the advancement of deep learning and have been widely applied in various computer vision tasks. Lu et al. [26] introduced an adaptive attention model with a visual sentinel tailored for image captioning. Fan et al. [27] utilized stacked latent attention layers to bolster multimodal reasoning capabilities. Yan et al. [28] developed a spatial attention mechanism within a modified convolution network to suppress irrelevant features during frame fusion. Similarly, Tel et al. [29] combined spatial attention with a semantic-consistent cross-frame attention block, enhancing the understanding of dynamic and static image content. In addition, Tao et al. [30] applied an adaptive interference removal framework in video person re-identification, with

attention-guided modules to selectively remove frame and pixel interferences. Abdusalomov et al. [31] employed spatial and channel attention synergistically to improve the accuracy of detection of brain tumors from MRI scans. Tao et al. [32] made strides in forest smoke detection by developing an attention-steered, pixel-level supervision model that differentiates subtle features effectively. Liu et al. [33] introduced an innovative pyramid cross-attention alignment module, which efficiently aggregates features from LDR frames, achieving both denoising and HDR reconstruction. Moving beyond traditional spatial attention, several fusion network architectures with expansive receptive fields, such as non-local networks [34] and the transformative transformer networks [18], have been proposed. By focusing on pertinent information, these networks set new performance standards, demonstrating increased adaptability. The proposed method uniquely integrates attention guidance to accentuate this capability.

2.3. Vison Transformer

The vision transformer [16] (ViT) model, initially designed for NLP tasks, has been adapted for visual tasks. Dosovitskiy et al. [35] were at the forefront of introducing the ViT for image recognition, segmenting images into input tokens and analyzing their interrelationships. However, ViT's performance aligns with CNNs only when subjected to extensive dataset training. Touvron et al. [36] refined ViT using knowledge distillation, emulating the output of a CNN model teacher. To mitigate the self-attention computational overhead in high-resolution images, Liu et al. [37] proposed a hierarchical transformer structure with an integrated moving window. However, this method might curtail the contextual information within local regions, rendering it potentially unsuitable for HDR reconstruction. To address this, we reshape the image patches within the transformer and incorporate convolution operations, ensuring the comprehensive capture of both local and global information. The proposed method is benchmarked against leading-edge CNN methods, underscoring its effectiveness and feasibility.

3. Proposed Single-Image HDR Reconstruction Framework

The comprehensive structure of the HDRFormer framework is depicted in Figure 2. An encoder–decoder model, serving as the foundational architecture, is detailed in Section 3.1. This architecture leverages hierarchical multi-scale features, amalgamating low-level and high-level features. In Section 3.2, the pivotal roles of the weighted attention mechanism (WAM) and the modified specular-free (MSF) are introduced, both of which are paramount in accentuating saturated regions. Section 3.3 provides a detailed explanation of the loss functions employed for training the HDR model.

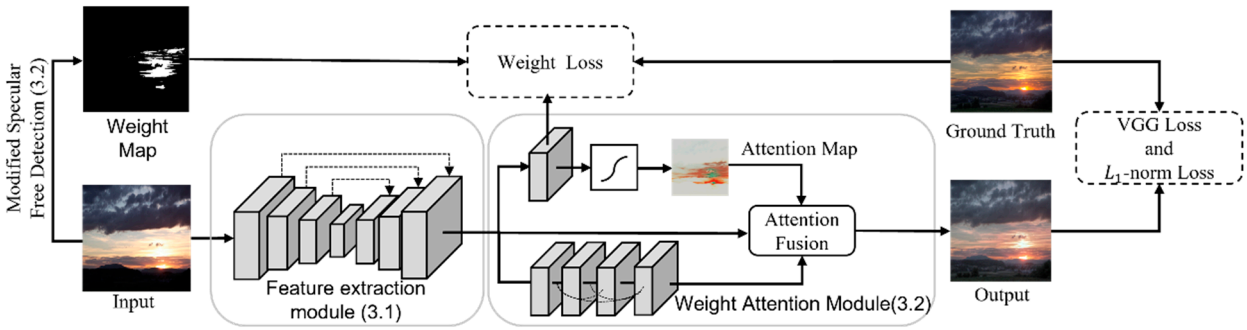


Figure 2. Overview of the proposed HDRFormer. The model begins with the input LDR image, which is processed through the FEM, aiming to reconstruct the information of the HDR. Subsequently, the MSF is employed to identify regions of interest, while the WAM is invoked to further refine these regions, ensuring optimal enhancement. Ultimately, the model is trained and the final HDR images are obtained through the calculation of loss.

3.1. Feature Extraction Module

The proposed feature extraction module is founded on a U-net-based encoder–decoder architecture.

Encoder. Initially, we extract shallow features, $F_0 \in \mathbb{R}^{H\times W\times C}$, from an input image $I \in \mathbb{R}^{H\times W\times 3}$ using 3×3 convolutional layers, where H, W, and C represent the image height, width, and the number of feature channels, respectively. For this module, we set the feature channels to 32 based on the experimental findings. The feature map F_0 is extracted via three encoder stages, each consisting of two LASwin transformer blocks, one edge-guided reinforcement block (applied only to stages 1 and 2), and a downsampling layer. The LASwin transformer block captures global context using a self-attention mechanism, while a feedforward network focuses on local details. The edge-guided reinforcement block utilizes guided filters to preserve edge nuances, subsequently enhancing the feature map's structural information via convolution operations. Downsampling is performed using a 4×4 convolutional layer with a stride of 2, reducing the image size by half while doubling the number of channels. Thus, the feature of the i-th stage in the encoder is defined as $F_{\mathrm{i}} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 2^i C}$, where $i \in \{1, 2, 3\}$ represents the three stages of the encoder. Concluding the encoder, we append two LASwin transformer blocks, forming a bottleneck stage, and produce the feature map F_3'.

Decoder. We have a symmetric three-stage design mirroring the encoder. Each stage is equipped with two LASwin transformer blocks, an edge-guided reinforcement block, and an upsampling layer. The feature of the i-th stage in the decoder is expressed as $F_i' \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 2^i C}$,where $i \in \{0, 1, 2\}$ represents the decoder's three stages. Upsampling is performed by merging bilinear interpolation with a 3×3 convolutional layer, reducing the channel count by half, and expanding the feature map dimensions. The resulting feature map is then merged with the corresponding encoder features. Skip connections are used to channel these features to the LASwin transformer block, facilitating HDR image reconstruction. The edge-guided reinforcement blocks ensure that the texture details of the HDR image are accurately rendered. Finally, the feature map F_0' is passed through the WAM to emphasize saturated regions. The following sections will provide a more in-depth understanding of the LASwin transformer block and the edge-guided reinforcement block.

LASwin transformer block. The transformer architecture has emerged as a potential solution to the challenges encountered by CNN-based methods, especially in capturing long-range dependencies. However, experiments have indicated that the direct integration of the transformer into HDR reconstruction is fraught with challenges. While the transformer is renowned for its robust learning capabilities, achieving convergence remains a hurdle. Consequently, we employ handcrafted and masking techniques to guide feature orientations, enhancing convergence and performance. Moreover, transformers face two primary challenges. Firstly, while segmenting images into uniform patches and computing global self-attention across them is feasible, the computational demands of the transformer increase quadratically with the number of patches [37]. This can be taxing on hardware resources, sometimes to the point of being prohibitive. Therefore, applying global self-attention to high-resolution feature maps is not always practical. Secondly, as indicated by prior research [38,39], transformers tend to be less proficient in capturing local information, often overlooking the internal structural details, especially when compared to CNNs. To overcome these limitations, we introduce the local-attention shifted windows (LASwin) transformer block, depicted in Figure 3a. This block leverages the transformer's self-attention mechanism to capture global dependencies and incorporates a feedforward network to capture salient local feature nuances. Specifically, the LASwin transformer block comprises two main modules: (1) window-based multi-head self-attention (W-MSA) and (2) local attention feedforward network (LAFF). The computational dynamics of the LASwin transformer block are detailed as follows:

$$F' = \mathrm{W} - \mathrm{MSA}(\mathrm{LN}(F_{in})) + F_{in}, \tag{1}$$

$$F_{out} = \mathrm{LAFF}\big(\mathrm{LN}\big(F'\big)\big) + F', \tag{2}$$

where F_{in} represents the input feature map of W-MSA, and LN $(\cdot)$ denotes the layer normalization. F' and F_{out} are the output features of W-MSA and LAFF, respectively. We adopt the window-based multi-head self-attention [37], in contrast to the global multi-head self-attention (MSA) typical of the standard transformer. By performing self-attention within distinct, non-overlapping local windows, the computational complexity scales linearly with the spatial size, leading to a significant reduction in the computational cost. Specifically, for an input feature map F_{in} of size $H \times W \times C$, the LASwin transformer block divides the input into non-overlapping $M \times M$ local windows, where $M = 4$. These windows are subsequently flattened and transposed, resulting in $\frac{HW}{M^2} \times M^2 \times C$ features, where $\frac{HW}{M^2}$ represents the total number of windows. Following this, each local window's self-attention is computed. For the i-th window feature input $F_{in}^i \in \mathbb{R}^{M^2 \times C}$, the query Q^i, key K^i and value V^i are derived as follows:

$$Q^i = F^i P_Q, \;\; K^i = F^i P_K, \;\; V^i = F^i P_V, \tag{3}$$

where $i \in \left\{1, 2, \ldots, \frac{HW}{M^2}\right\}$, and P_Q, P_K, and P_V are learnable parameters. We divide Q^i, K^i, and V^i into n heads across the channel dimension, with each head having a size defined by $d_{size} = \frac{HW}{n}$. The self-attention of the j-th head is expressed as

$$\mathrm{Attention}\left(Q_j^i, K_j^i, V_j^i\right) = \mathrm{SoftMax}\left(\frac{Q_j^i {K_j^i}^T}{\sqrt{d_{size}}}\right) V_j^i, \tag{4}$$

where $j \in \{1, 2, \ldots, n\}$, and Q_j^i, K_j^i, and V_j^i represent the query, key, and value of the j-th head of the i-th window feature, respectively. The output for the i-th window feature, $F_o^i \in \mathbb{R}^{M^2 \times \mathrm{C}}$, can be formulated as

$$F_o^i = C_{j=1}^n \left(\mathrm{Attention}\left(Q_j^i, K_j^i, V_j^i\right)\right) P_O + B, \tag{5}$$

where $C(\cdot)$ denotes the concatenation operation, $P_O \in \mathbb{R}^{\mathrm{C} \times \mathrm{C}}$ is a learnable parameter, and $B \in \mathbb{R}^{M^2 \times \mathrm{C}}$ is a learnable relative position embedding. Subsequently, all window features are reshaped to derive the output feature map F'. However, since the segmented local windows do not overlap, this might result in a lack of information interaction between windows. To address this, we alternate between standard windows and shifted windows, ensuring cross-window connections [37]. Specifically, shifted windows involve moving the original feature data upwards and to the left by half the window size before implementing the standard self-attention mechanism.

On the other hand, the feedforward network (FFN) in the standard transformer processes each pixel location independently and uniformly for feature transformation. Specifically, FFN employs two 1×1 convolutional layers: the first expands the feature channels, while the second reverts the channels to their original input dimensions. In addition, a non-linear function, GeLU, is applied within the hidden layer. However, as prior research [38] pointed out, FFNs exhibit limitations in handling local features. The contribution of adjacent pixels is important for image reconstruction [40]. To address this, we have adapted the FFN within the transformer-based framework by integrating a depthwise convolution block and a squeeze-and-excitation (SE) block [41]. As shown in Figure 3b, the proposed approach begins with applying a linear projection layer (1×1 convolution) to each input token, enhancing the feature dimension and reshaping it into a 2D feature map. Subsequently, a 3×3 depthwise convolutional layer is employed to extract the image's local nuances, which is crucial for image reconstruction. The SE block is then utilized for channel weighting, with the GeLU function activated in the hidden layer.

This channel weighting strategy facilitates the transmission of relevant information while filtering unnecessary data, mirroring the attention mechanism's functionality. Finally, we flatten the feature map and apply another linear layer (1×1 convolution) to reduce the channels and align them with the input channel dimensions, followed by a residual output.

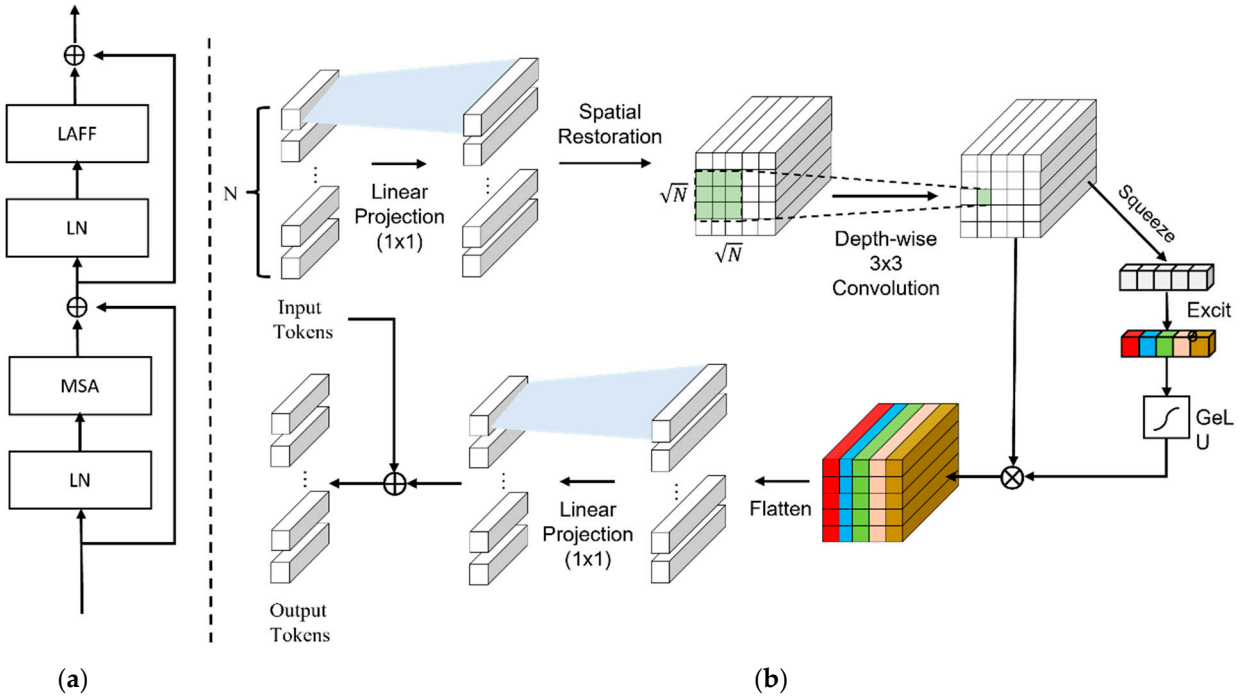


Figure 3. (**a**) Illustrates the detailed structure of the LASwin transformer block, utilizing a window-based multi-head self-attention mechanism to capture distant dependencies and employing a local attention feedforward network (LAFF) to learn local information. (**b**) showcases the LAFF. Initially, input patch tokens are projected into a higher dimension and subsequently restored to a 2D image in the spatial dimension, adhering to the original position. A deep convolution, indicated by the green area, is then performed on the restored image. Finally, channel weighting is applied to the image, which is then flattened and projected to the initial dimension for residual output.

Edge-guided reinforcement block. The encoder–decoder's upsampling operation can often compromise edge details during image compression and reconstruction. Although U-net can improve the image's edge information via skip connections, there is an inevitable loss of some edge structural details. Drawing inspiration from [42], we employ guided filters after downsampling and before upsampling to retain and amplify the image's edge structural details. Recognized for their edge-preserving capabilities, guided filters have been widely adopted in traditional handcrafted methods for tasks such as image enhancement, dehazing, and HDR compression. In the context of deep learning, the linear nature of the guided filter's computational process allows it to serve as a layer, facilitating gradient computation and subsequent backpropagation.

Algorithm 1 outlines the procedure for edge-guided reinforcement blocks (ERBs). The ERB integrates a guided filter and a 3×3 convolutional layer, with the feature map F and the guided image G as inputs. We employ a gray-scale image as the guide image, aiming to retain the guide image's edge structural details through guided filtering applied to the feature map. The process begins with a mean filter f_{mean} with a radius of 2 to compute the correlation coefficient. This coefficient encompasses the mean of the feature map μ_F, the mean of the guide image μ_G, the variance of the feature map σ_G^2, and the covariance between the feature map and guidance image $\sigma_{G,F}$. Using the derived variance σ_G^2 and covariance $\sigma_{G,F}$, these coefficients collectively discern the linear relationship between the feature map and the guide image. Subsequently, the linear transformation parameters a and b are deduced using the least squares method. To prevent the parameter a from

becoming excessively large, we introduce the parameter ε, set to $\varepsilon = 10^4$ based on our experiments. The final step involves calculating the mean of parameters a and b, using the derived mean parameter to linearly generate the output feature map O for the fine structure. However, while the guided filter can extract the guided image's fine structural details, its application to the feature map might inadvertently smooth out certain flat feature details. Therefore, we concatenate the output map O with the original feature map F, using a 3×3 convolutional layer to further extract and preserve essential edge structural details, enhancing the model's performance.

Algorithm 1: Edge-guided reinforcement block

Input: Feature map F, guidance image G, radius r, regularization term ε
Output: Feature map F'
Step 1: Utilize box filter f_{mean} (.) to compute the correlation coefficient, which includes μ_F, μ_G, σ_G^2, and $\sigma_{G,F}$.

$$\mu_F = f_{\text{mean}}(F, r)$$
$$\mu_G = f_{\text{mean}}(G, r)$$
$$\sigma_G^2 = f_{\text{mean}}(G^2, r) - \mu_G^2$$
$$\sigma_{G,F} = f_{\text{mean}}(G \cdot F, r) - \mu_G \cdot \mu_F$$

Step 2: Determine the optimal linear transformation parameter coefficients a and b based on the given formulas.

$$a = \frac{\sigma_{G,F}}{\sigma_G^2 + \varepsilon}$$
$$b = \mu_F - a \cdot \mu_G$$

Step 3: Compute the mean values of parameters a and b using the provided formulas and derive the output image O using guided filtering.

$$\mu_a = f_{mean}(a, r),\ \mu_b = f_{mean}(b, r)$$
$$O = \mu_a \cdot F + \mu_b$$

Step 4: Concatenate F with O and perform a convolution to fuse the feature and modify the channel size of the feature.

$$F' = Cov(Concat(F, O))$$

3.2. Weight Attention Module

One of the most significant challenges in HDR image reconstruction is the restoration of details in saturated regions. The lack of these details can significantly compromise the visual quality of the image. The learning efficacy of a neural network model is substantially influenced by the appropriate selection of architecture and loss functions. As observed in previous studies [43,44], many existing approaches focus solely on minimizing the difference between the reconstructed and target images, often neglecting the crucial aspect of restoring saturated regions. To address this issue, we introduce a weight attention module (WAM) designed to enhance the information in saturated regions, thereby improving model performance. As illustrated in Figure 4, the WAM offers two main contributions. First, it provides precise guidance for the saturated regions. Second, it employs saturation supervision to generate attention maps, which are pivotal in enhancing the details within the saturated regions while concurrently minimizing errors in other areas. To precisely target saturated regions, we first identify these areas within the image and generate a corresponding weight map. While traditional methods detect saturated regions quickly, employing fixed thresholds may not be universally applicable. In contrast, we use the modified specular free (MSF) method [45] for the adaptive detection of saturated regions, thereby achieving self-supervised learning for these regions. The MSF method determines saturated regions based on the difference between the original input image I and the MSF image. Given the high-intensity variance in overexposed regions, we primarily focus on their restoration. Upon identifying saturated regions using MSF and generating the weight map, the weight map is calculated as follows:

$$W(i,j) \in \begin{cases} \frac{\max\left(0, \max\left(\delta_x^R(i,j),\ \delta_x^G(i,j), \delta_x^B(i,j)\right) - th\right)}{(255 - th)}, & \text{if } \delta_x(i,j) > th \text{ for all } x \\ 0, & others \end{cases}, \tag{6}$$

where $W(i,j)$ represents the weight map, $\delta_x(i,j) = I(i,j) - MSF_x(i,j)$, and th is the threshold calculated using the Otsu algorithm. We then feed the feature maps generated by the base network into a 3×3 convolution layer to produce the predicted HDR image. For this predicted HDR image, we compute a weighted loss using the weight map W. This allows us to focus on the saturated regions, generating an attention mask M through a 1×1 convolution layer followed by a sigmoid activation function. On the other hand, we employ three 3×3 convolution layers for dense connections to learn the image features and utilize a 1×1 convolution layer and two 3×3 convolution layers for the attention mechanism. The attention mechanism is primarily achieved through residual learning and the attention mask M, ultimately outputting higher-quality HDR images. The underlying rationale for this design is to provide additional guidance to the model in saturated regions through the self-training of attention.

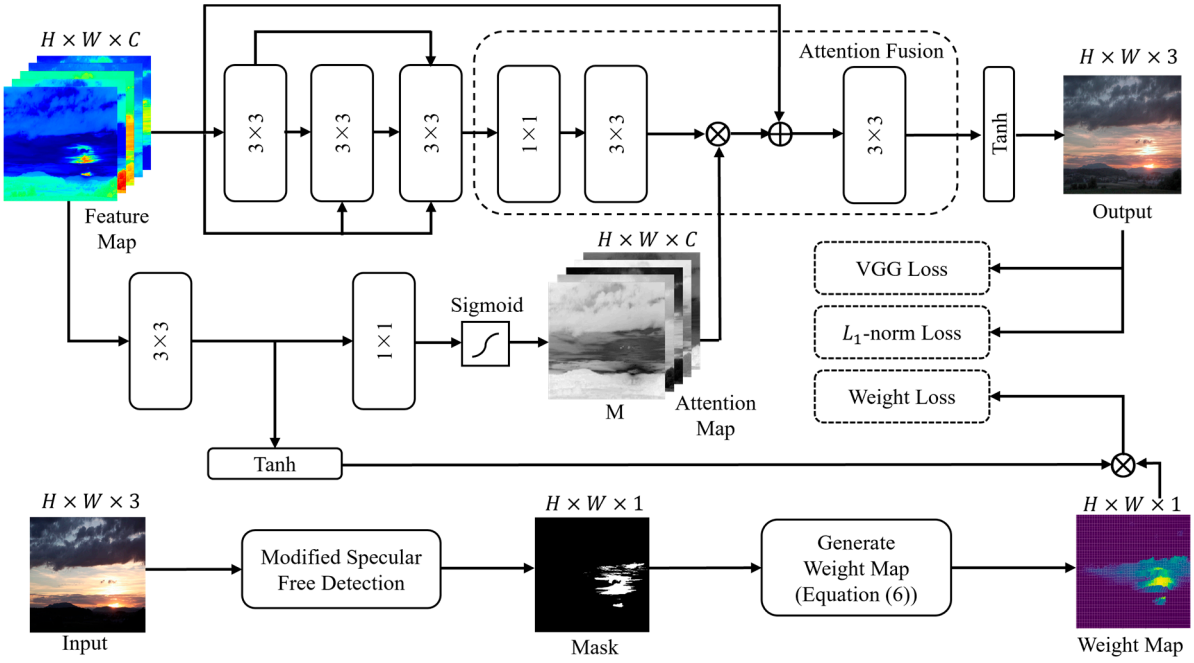


Figure 4. Detailed architecture of the WAM. Initially, to accurately enhance the information in the saturated regions, we employ the MSF detection method to identify the locations of the saturated areas and generate a weight map accordingly. Subsequently, the feature map anticipates HDR reconstruction information through convolution operations and integrates the preceding weight map to focus on the restoration of saturated areas, subsequently transforming this information into an attention map. Thirdly, the feature map undergoes dense convolution operations to learn the HDR reconstruction information and merges with the generated attention map for attentive fusion output. Finally, loss calculation is performed on the output image.

3.3. Loss Function

In various image tasks, researchers typically design suitable loss functions according to actual needs to ensure that the model converges in the desired direction. For deep learning-based HDR image reconstruction methods, considering that HDR images are commonly displayed post-tone mapping, we compute the loss function between the tone-mapped generated HDR image and the tone-mapped ground truth HDR image. We adopt the μ-law for tone mapping, with the calculation formula as follows:

$$T(H) = \frac{log(1+\mu H)}{log(1+\mu)}, \tag{7}$$

where $T(H)$ denotes the operation of HDR image H with μ-law, and μ defines the amount of compression. Following the settings in reference [10], we set $\mu = 5000$ and keep the

image range within [0, 1]. Next, we calculate all loss functions between the tone-mapped generated HDR image and the tone-mapped HDR image of ground truth. Specifically, we utilize a combination of three loss functions to train the model. Equation (8) describes our final loss function, which is the weighted sum of reconstruction loss L_{Rec}, weight loss L_w, and perceptual loss L_p. In this work, we set $\lambda = 10$.

$$L_{total} = L_{Rec} + L_w + \lambda L_p. \tag{8}$$

Reconstruction loss. The objective of model training is to generate output HDR images that closely resemble the HDR image of the ground truth. Therefore, we adopt the pixel-level L_1-norm loss function as our HDR reconstruction loss. Compared to L_2-norm, L_1-norm yields fewer blurry results and is robust to outlier pixel values. However, the range of linear HDR brightness values can span from very low (shadow areas) to very high (bright areas). The direct application of the reconstruction loss function on linear HDR images may not produce optimal results. From experience, we found that calculating the loss on images generated using tone mapping results in more stable training and better performance. Therefore, we use the tone mapping operator to map the generated HDR image and the HDR image of ground truth separately and calculate their reconstruction loss. The reconstruction loss function calculation formula is as follows:

$$L_{\text{Rec}} = \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{3} \left(T\left(H_{gen}^{(h,w,c)} \right) - T\left(H_{\text{gt}}^{(h,w,c)} \right) \right). \tag{9}$$

where H and W are the image height and width, respectively, H_{gen} represents the generated HDR image, and H_{gt} represents the HDR image of the ground truth. $T(\cdot)$ is the μ-law tone mapping operator.

Perceptual loss. To generate more reasonable and realistic details in the output HDR image, we use perceptual loss [46]. This loss function attempts to evaluate the match degree of the reconstructed image features with the features extracted from the ground truth, enabling the model to produce a feature representation akin to the ground truth. The perceptual loss function calculation formula is as follows:

$$L_p = \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{3} \left(\phi_l \left(H_{\text{gen}}^{(h,w,c)} \right) - \phi_l \left(H_{\text{gt}}^{(h,w,c)} \right) \right), \tag{10}$$

where $\phi_l(\cdot)$ represents the feature map output extracted from the lth layer of the VGG-19 network. Here, the L_1-norm loss between the feature maps of the tone-mapped generated HDR image and the tone-mapped HDR image of the ground truth is calculated.

Weight loss. To generate higher quality HDR images in the output saturation areas, we designed a weight loss, as detailed in Section 3.2. This loss function identifies the location of saturated areas in the image and generates a weight map, enabling the network to focus on evaluating values in pixel-level saturated areas during training. The weight loss function calculation formula is as follows:

$$L_w = \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{3} W^{(h,w)} \left(T\left(H_{\text{pre}}^{(h,w,c)} \right) - T\left(H_{\text{gt}}^{(h,w,c)} \right) \right), \tag{11}$$

where W represents the weight values of the saturated areas detected by MSF, and H_{pre} represents the HDR image predicted earlier in the weighted attention module. Here, the loss of the pixel in the saturated area between the tone-mapped predicted HDR image and the ground truth HDR image is calculated.

4. Experiment

In this section, we initially delineate the experimental setup, which encompasses the datasets and training phase. Subsequently, we compare it quantitatively and qualitatively with other single-image HDR reconstruction methods. Finally, an ablation study is described to evaluate the contribution of each component.

4.1. Datasets and Experiment Settings

The selection of datasets plays a crucial role in determining the effectiveness of a deep learning network, facilitating learning, and maintaining objectivity in environments with similar objects. To robustly evaluate the merits of the proposed method, we employed widely recognized HDR image reconstruction datasets: Funt [47], Stanford [48], Ward [49], and Fairchild [50]. Given that these datasets solely encompass HDR images and lack LDR–HDR pairs, we applied tone mapping to synthesize corresponding LDR images, thereby obtaining a total of 9870 LDR images by randomly cropping 512 × 512 regions ten times for each tone mapping, in alignment with the methods delineated in [24,51,52]. In addition, we utilized the HDR-Real dataset [12], which comprises 480 HDR and 4893 LDR images. The LDR images, captured with 42 cameras, were converted into HDR images using Photomatix software version 6.3 [53]. Consistent with the approaches in [12,24,51,52], we amalgamated these datasets, referring to them as the HDR-Synth-Real dataset, and conducted the evaluations. A total of 11,810 images were allocated for training and 2953 images were allocated for testing. Furthermore, we conducted tests using the HDR-eye [54] public dataset, which includes 46 sets of ground truth HDR images and LDR images captured with various cameras. However, due to the substantial black areas present in the LDR images of the initial four sets, only 42 sets of the HDR-eye dataset were utilized for the evaluation.

Regarding the experimental environment settings, training and testing were implemented under the PyTorch framework on an Intel i7-8700K CPU equipped with an Nvidia GTX 1080 Ti GPU. The training images were resized to 256 × 256 pixels, and the HDR images were normalized to the [0, 1] range to ensure uniform learning. For testing, we utilized images of 512 × 512 pixels. The training was conducted over 200 epochs, with the interim results assessed every ten epochs to avert overfitting. In terms of parameters, the Adam optimizer was configured with $\beta 1 = 0.9$, $\beta 2 = 0.999$, and a learning rate of 0.001, adopting a linear decay strategy every 30 epochs. The batch size was established at 1.

4.2. Qualitative Comparison

In the ensuing visual analysis, Photomatix [53] is employed for tone mapping, utilizing the 8-bit LDR image for the HDR image to compare with the proposed method, aligning with the prevalent practices in the majority of HDR literature. We juxtapose our approach with the recent handcrafted-based methods rTMO [55] and MEO [56], and the deep learning-based methods reported by Liu et al. [12], HDRCNN [13], ExpandNet [24], FHDR [25], and KUNet [57]. Visual comparisons enable the discernment of intuitive performance advantages among various methods, such as the detail retention in image regions, naturalness, and saturation levels, as exemplified in Figure 5a, 6a, and 7a. These test images, which exhibit significant luminance disparities in highlight or shadow regions, clearly reveal numerous detail losses upon visual inspection.

Figure 5 illustrates the results of reconstructing test image 07 from the HDR-eye dataset using different methods. In Figure 5c, although the method [55] enhances the overall brightness of the image, it exhibits a poor detail performance in the shadow (blue box) region, rendering it indistinct to the human eye. Conversely, Figure 5d retains some detail in the shadow (blue box) region, yet the arch details in the bright (red box) region lack clarity. In Figure 5e, the shadow (blue box) region details are less prominent, possibly due to the HDRCNN [13] neglecting the underexposed information. Figure 5f has a commendable overall naturalness, yet the tree bark texture details in the shadow (blue box) region are nearly imperceptible. Figure 5g represents shadow (blue box) and bright (red box) region details adequately, albeit with a somewhat unnatural overall color tone. Figure 5h successfully retains details in both the shadow (blue box) and light (red box) regions, but it presents a muted color vibrancy and insufficient global contrast. Figure 5i displays enhanced color saturation in both shadow (blue box) and light (red box) regions; however, it does not effectively capture the finer details of the architectural features and tree bark textures. In contrast, Figure 5j adeptly reconstructs the details in the shadow (blue

box) and bright (red box) regions, accurately reflecting the information of the real scene and exhibiting excellent natural color tones.

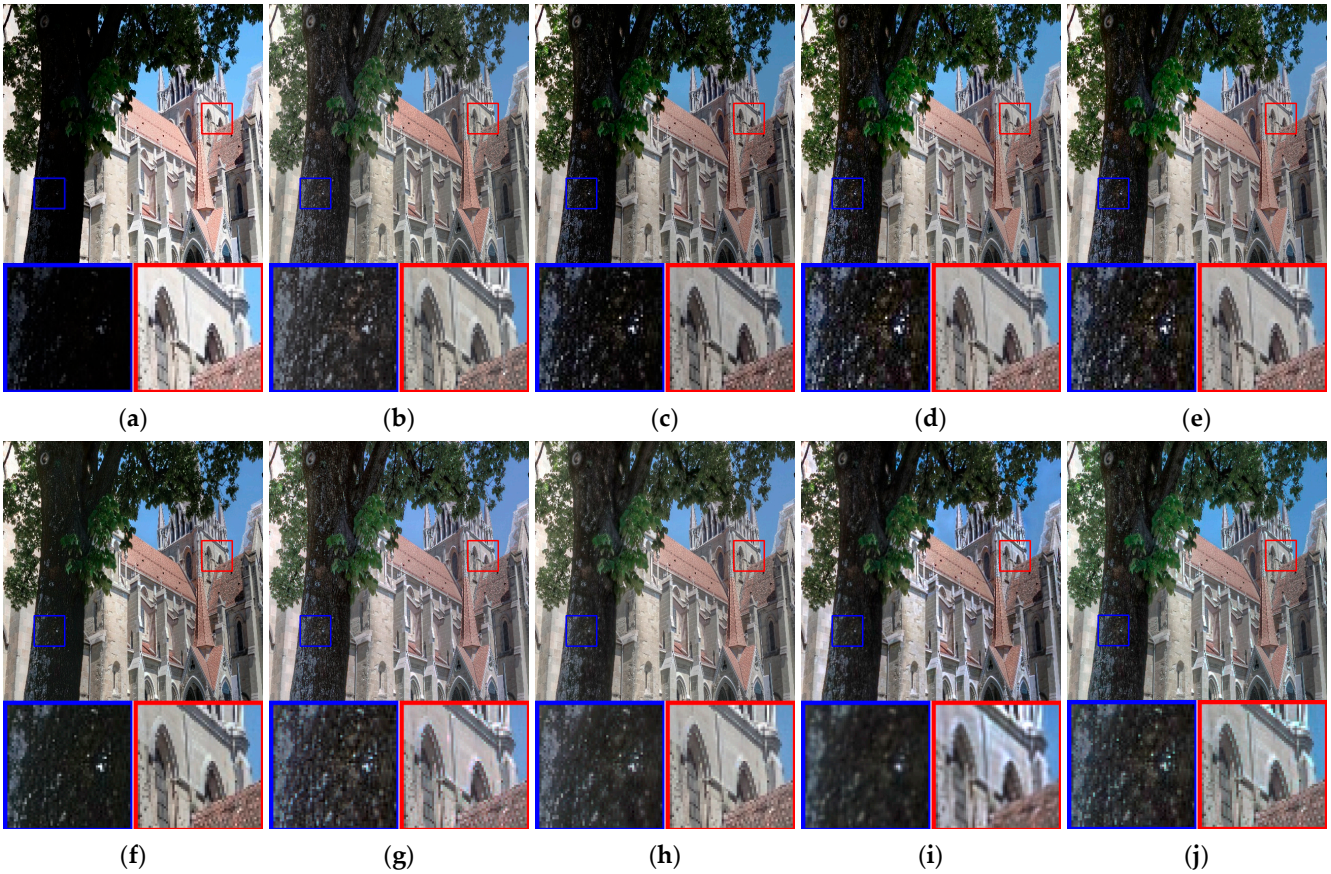


Figure 5. Comparative visualization of HDR image reconstruction applied to test image 07 from the HDR-Eye dataset. (**a**) LDR image, (**b**) ground truth, (**c**) rTMO [55], (**d**) MEO [56], (**e**) HDRCNN [13], (**f**) ExpandNet [24], (**g**) FHDR [25], (**h**) Liu et al. [12], (**i**) KUNet [57], and (**j**) the proposed method.

Figure 6 demonstrates the visualization of reconstructing test image 17 in the HDR-eye dataset. In Figure 6c, the stone texture details near the bright (red box) region are evident, yet the stair details in the shadow (blue box) region are not recovered. Figure 6d reveals texture details in the shadow (blue box) region, but the stone step details in the bright (red box) region are elusive. In Figure 6e, while the bright (red box) region details are preserved, noise is generated in the stairs in the shadow (blue box) region. Figure 6f displays details of the scene in the bright (red box) and shadow (blue box) regions, albeit with a monotonous and flat overall color tone. Figure 6g reconstructs details well in the bright (red box) and shadow (blue box) regions, maintaining good overall naturalness, yet the overall lack of contrast yields a suboptimal visual experience. Figure 6h improves the scene's visibility in both regions; however, the stair structure in the shadowed region (blue box) lacks clarity. Figure 6i more accurately represents the shapes in the bright area (red box) but appears somewhat blurred compared to the alternative approaches, presenting a color tone and contrast that deviate significantly from the ground truth. In comparison, Figure 6j, which benefits from the utilization of the ERB module, adeptly preserves the image's edge information, rendering the stair details in the shadow (blue box) region clearly visible, and the image overall is vibrant, presenting good visual richness.

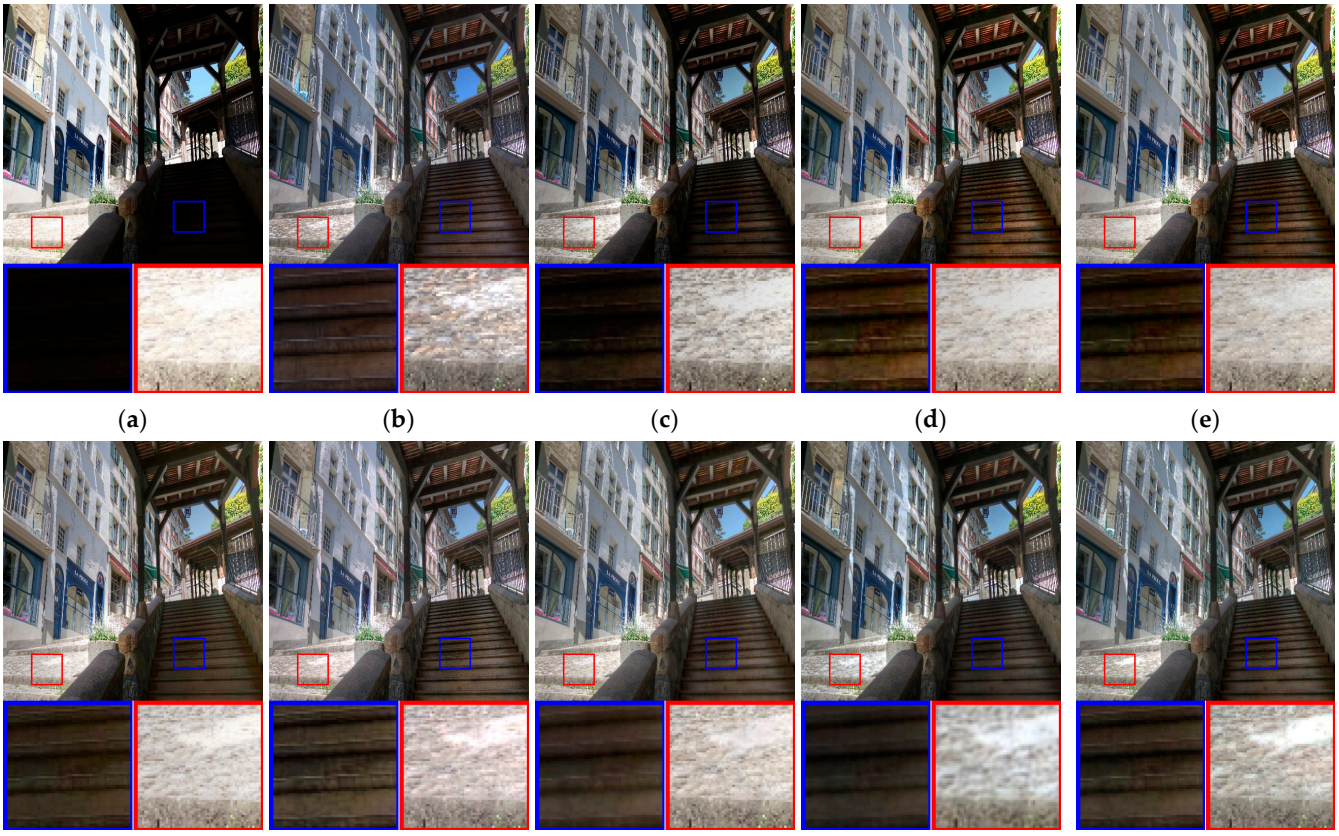


Figure 6. Comparative visualization of HDR image reconstruction applied to test image 17 from the HDR-Eye dataset. (**a**) LDR image, (**b**) ground truth, (**c**) rTMO [55], (**d**) MEO [56], (**e**) HDRCNN [13], (**f**) ExpandNet [24], (**g**) FHDR [25], (**h**) Liu et al. [12], (**i**) KUNet [57], and (**j**) the proposed method.

Figure 7 displays the visualization of reconstructing test image 22 in the HDR-eye dataset. In Figure 7c,d, the scene details in the bright (red box) and shadow (blue box) regions are almost invisible. Figure 7e preserves the bright (red box) region details but fails to effectively reconstruct the tile tones. Figure 7f displays the bright (red box) region details and maintains a natural overall color tone; however, the details in the shadow (blue box) region are lost. Figure 7g preserves details well, yet the overall color tone is monotonous, lacking contrast. Figure 7h represents the bright (red box) and shadow (blue box) region details well, yet the image appears unrealistic due to a lack of global contrast. While Figure 7i falls short in recovering additional structural details in the shadow (blue box) areas, the incorporation of the cascade structure from KUNet [57] facilitates the reconstruction of intricate structural details in the stairway section. In contrast, Figure 7j reconstructs more visual content in the bright (red box) and shadow (blue box) region details, thereby enhancing the overall visual quality.

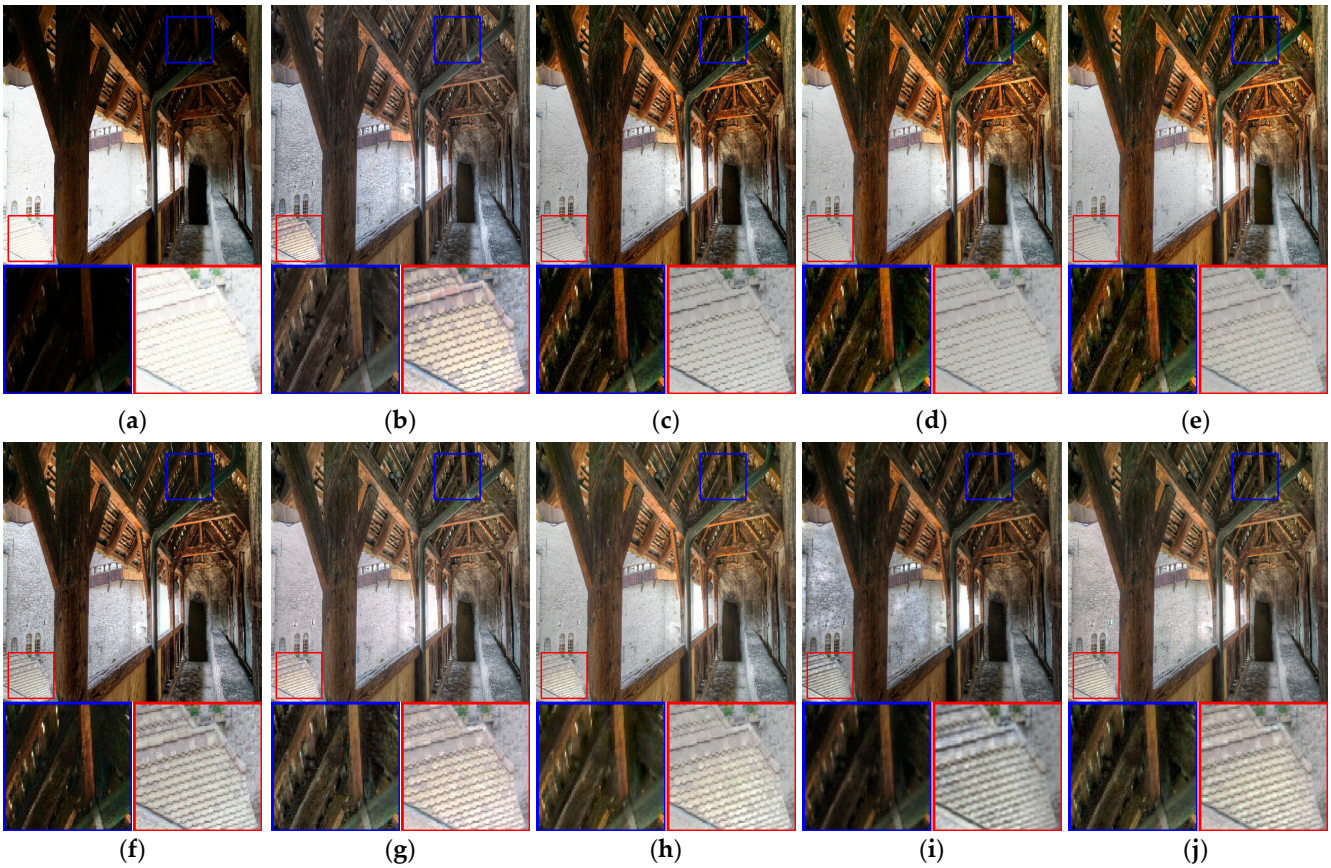


Figure 7. Comparative visualization of HDR image reconstruction applied to test image 22 from the HDR-Eye dataset. (**a**) LDR image, (**b**) ground truth, (**c**) rTMO [55], (**d**) MEO [56], (**e**) HDRCNN [13], (**f**) ExpandNet [24], (**g**) FHDR [25], (**h**) Liu et al. [12], (**i**) KUNet [57], and (**j**) the proposed method.

4.3. Quantitative Comparison

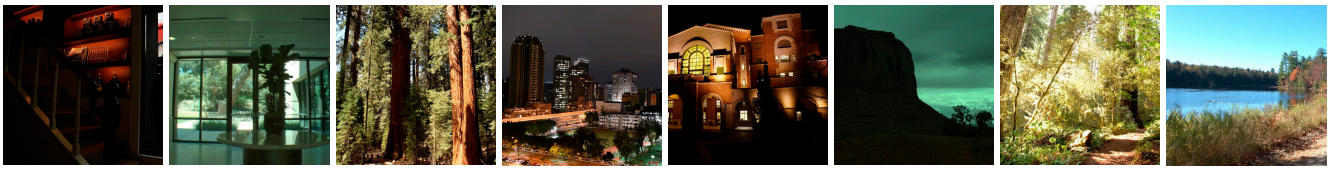
To ensure a fair comparison, we utilized commonly used metrics such as the peak signal-to-noise ratio (PSNR) [58], structural similarity (SSIM) [58], and multi-scale structural similarity (MS-SSIM) [59] to evaluate the differences between the predicted HDR images and the HDR ground truth images, thereby objectively comparing the advantages among the different methods. The PSNR is calculated based on the mean squared error (MSE), which represents the Euclidean distance between pixels and indicates a higher similarity to the original pixels as its value increases. The SSIM measures the similarity to the original image using pixel variance and the mean of two images, while the MS-SSIM incorporates scale-space theory into the SSIM calculations. Before calculating the PSNR, SSIM, and MS-SSIM metrics, we applied the μ-law tone mapping operator to the images, maintaining the output within [0, 1]. Additionally, HDR images should be measured based on human visual perception and image differences. Therefore, we also employed the quality Q score of the HDR visual difference predictor (HDR-VDP2.2) [60] as an evaluation metric, which is based on a human visual perception model and is applicable to all luminance environments, providing the highest level of objective evaluation of the HDR image quality.

Table 1 displays the comparison of the various methods' PSNR, SSIM, MS-SSIM, and HDR-VDP2.2 scores on two datasets. Figure 8 shows the thumbnails of test image scenes, such as outdoor/indoor, day/night, and rural/urban scenarios. On the test dataset, deep learning-based methods outperformed traditional inverse tone mapping algorithms in average scores across all metrics, as it might be challenging for traditional methods to find optimal parameters and functions for each image. In contrast, deep learning-based methods, with superior adaptability and minimal training parameter tuning, demonstrate robust HDR reconstruction capabilities across various images. Notably, the proposed method

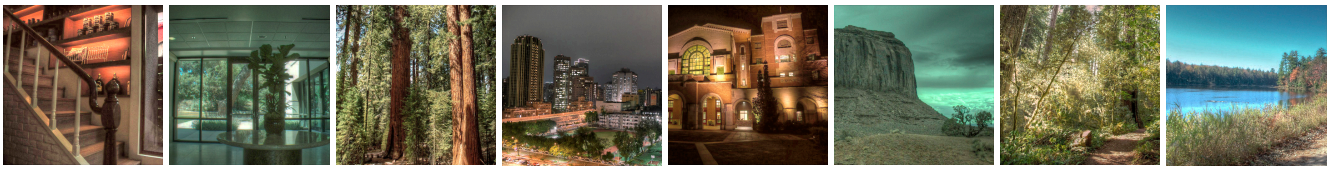
yielded the best PSNR, MS-SSIM, and HDR-VDP2.2 scores and the second-best SSIM score. A higher PSNR score indicates lower distortion in the generated image, implying a better quality of the predicted image. A higher HDR-VDP2.2 score indicates a better quality of the generated image, meaning the predicted image is closer to the real image under human visual system observation. The SSIM and MS-SSIM scores denote the structural similarity of the generated image, with higher scores indicating a greater structural similarity between the predicted and real images. Although our method slightly underperformed in the SSIM metric compared to [25], considering the scale concept, we achieved the best score in the MS-SSIM metric. Simultaneously, we also conducted evaluation tests on the publicly available HDR-eye dataset, and the proposed method achieved the best scores in all four metrics. Overall, compared to other HDR reconstruction methods, the proposed method achieves superior performance and exhibits a better generalization capability.

Table 1. Reconstruction results obtained for the HDR-Synth-Real and HDR-EYE public datasets. The bold indicates the best performance.

Method	HDR-Synth-Real				HDR-EYE			
	PSNR	**SSIM**	**MS-SSIM**	**Q-Score**	**PSNR**	**SSIM**	**MS-SSIM**	**Q-Score**
MEO [56]	14.3587	0.5558	0.7273	47.9059	15.3920	0.6789	0.8143	50.7127
rTMO [55]	15.7702	0.5903	0.7906	49.9514	15.6703	0.6932	0.8492	50.9251
HDRCNN [13]	15.8559	0.5893	0.7982	51.1234	16.9791	0.7024	0.8520	52.2274
ExpandNet [24]	15.7862	0.6122	0.7877	50.0693	17.4671	0.7555	0.8753	52.4345
Liu et al. [12]	16.8019	0.6322	0.8123	52.5716	19.5234	0.7999	0.9255	54.7989
FHDR [25]	17.1186	**0.6542**	0.8308	52.1636	19.4047	0.7913	0.9307	53.7691
KUNet [57]	17.2050	0.6206	0.8159	52.1499	18.9806	0.7531	0.9202	53.6063
This work	**17.4261**	0.6503	**0.8376**	**52.7729**	**19.5814**	**0.8105**	**0.9367**	**54.9493**



(**a**)



(**b**)

Figure 8. Thumbnails of test images. (**a**) LDR image, and (**b**) reconstruction result of the proposed method.

4.4. Ablation Studies

This section delineates the evaluation of contributions from various components within the proposed HDRFormer model. Specifically, we utilize LDR images as input and employ a CNN-based U-net model as the baseline [13]. The different components are adjusted and incorporated to analyze and evaluate their efficacy within the HDRFormer. Table 2 presents the performance evaluation of each component, articulated in terms of the PSNR, SSIM, and MS-SSIM. Initially, the advantages of the FEM are analyzed. A conspicuous improvement is observed in the PSNR, SSIM, and MS-SSIM evaluations, with values approximately 4.730, 0.069, and 0.064, respectively, upon substituting the CNN-based feature extraction module with the proposed LASwin transformer block. This result underscores the aptness of the LASwin transformer block architecture for HDR reconstruction.

Table 2. Results of ablation studies using the FEM and WAM on the HDR-Synth-Real dataset.

FEM		WAM	PSNR	SSIM	MS-SSIM	Runtime (ms)
LASwin Transformer Block	ERB					
Unet backbone (Baseline) [13]			12.2037	0.5753	0.7675	205
✓			16.9339	0.6443	0.8370	322
✓	✓		17.1641	0.6490	0.8404	343
✓		✓	17.0415	0.6451	0.8370	381
✓	✓	✓	17.4261	0.6503	0.8376	395

The subsequent analysis delves into the disparities engendered by the incorporation of the ERB. The quantitative experimental results demonstrate improvements across all the evaluated metrics. Figure 9 visually elucidates the impact of embedding the ERB module within the LASwin transformer block architecture. In the absence of the ERB, the texture details of the wall, when illuminated by the streetlight, are somewhat obfuscated. In contrast, employing the ERB not only renders the texture details distinctly visible but also preserves the structural information of the wall. Furthermore, the efficacy of the WAM is assessed. While integrating the WAM into the LASwin transformer block architecture does not universally enhance the quantitative results, the visual analysis reveals a more natural reconstruction in saturated areas. Figure 10 illustrates that, without the WAM, artifacts appear in the sky region. Conversely, employing the WAM presents a natural scenery devoid of artifacts in the sky region. This result can be attributed to our strategy of focusing on the information in the saturated areas when utilizing the WAM, thereby providing it with a more accurate direction to reconstruct images in real-world scenarios.

(**a**) (**b**)

Figure 9. Comparison of HDR images with and without ERB. (**a**) Without ERB, and (**b**) with ERB.

In addition to component evaluation, a runtime analysis was conducted to provide insights into the computational efficiency of this study. Table 2 demonstrates that incorporating the LASwin transformer block, ERB, and WAM increases the computational runtime; processing an image takes approximately 395 ms with the complete HDRFormer model, in contrast to the 205 ms required by the baseline U-net model. Despite the increase, the significant improvements in image quality, as denoted by the superior PSNR, SSIM, and MS-SSIM metrics, warrant the additional computational cost. The comprehensive

HDRFormer model yields remarkably refined reconstruction results. These quantitative and visual enhancements in HDR images affirm the increased runtime, emphasizing the HDRFormer model's effectiveness in producing high-fidelity images within an acceptable processing time increment. In future work, we aim to optimize model inference techniques to enhance HDRFormer further, facilitating deployment on edge devices and applications in practical settings.

(**a**) (**b**)

Figure 10. Comparison of HDR images with and without WAM. (**a**) Without WAM, and (**b**) with WAM.

5. Conclusions

In this study, we propose a novel framework for HDR image reconstruction in edge–cloud computing environments, focusing on enhancing video surveillance systems. The framework leverages the self-attention mechanism inherent in transformers to capture long-range dependencies in images while incorporating convolutional operations for the more effective learning of local features. To further enhance the quality of reconstructed images, we integrated the ERB and the WAM, focusing on strengthening the information in edge structures and saturated regions, respectively. The comprehensive experimentation on publicly available HDR image reconstruction datasets has substantiated the exceptional performance of the HDRFormer, particularly evident in the qualitative comparison analysis, where the proposed method demonstrates superior overall appearance and detail restoration capabilities. The ablation studies further validated the efficacy of each component within the HDRFormer architecture.

Despite these advancements, we acknowledge certain limitations. The current HDRFormer, while robust, has yet to be fully optimized for computational efficiency, which is paramount for real-time applications. Furthermore, the framework's adaptability to different imaging modalities and potential applicability in real-world scenarios require additional exploration. To tackle these challenges, the future iterations of the HDRFormer will focus on enhancing the model's efficiency, reducing its computational cost, and refining model inference techniques to facilitate its deployment on edge devices. In addition, we intend to extend our examination of the framework's adaptability across imaging conditions and its broader applicability to real-world contexts.

Author Contributions: Y.-Y.C., C.-H.H., S.-Y.J. and C.-F.L. participated in the study design and drafted the manuscript. Conceptualization, Y.-Y.C. and S.-Y.J.; methodology, Y.-Y.C., C.-H.H. and S.-Y.J.; software, S.-Y.J.; validation, C.-H.H. and C.-F.L.; formal analysis, S.-Y.J.; investigation, S.-Y.J.; resources, Y.-Y.C. and C.-H.H.; data curation, S.-Y.J.; writing—original draft preparation, C.-H.H., S.-Y.J. and C.-F.L.; writing—review and editing, C.-H.H. and C.-F.L.; visualization, S.-Y.J.; project administration, Y.-Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: The support from the Intelligent Manufacturing Innovation Center (IMIC), National Taiwan University of Science and Technology (NTUST), Taipei 10607, Taiwan, which is a Featured Areas Research Center in Higher Education Sprout Project of Ministry of Education (MOE), Taiwan (since 2023) was appreciated.

Institutional Review Board Statement: Not applicable. This study did not involve any humans or animals.

Informed Consent Statement: Not applicable. This study did not involve humans.

Data Availability Statement: The data supporting the findings of this study are openly available in the HDR-Synth-Real and HDR-eye datasets repository, accessible at doi:10.1109/CVPR42600.2020.00172, reference number 12. In addition, the datasets can be found at https://alex04072000.github.io/SingleHDR/ (accessed on 7 November 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, Y.-Y.; Lin, Y.-H.; Hu, Y.-C.; Hsia, C.-H.; Lian, Y.-A.; Jhong, S.-Y. Distributed Real-Time Object Detection Based on Edge-Cloud Collaboration for Smart Video Surveillance Applications. *IEEE Access* **2022**, *10*, 93745–93759. [CrossRef]
2. Hsiao, S.-J.; Sung, W.-T. Intelligent Home Using Fuzzy Control Based on AIoT. *Comput. Syst. Sci. Eng.* **2023**, *45*, 1063–1081. [CrossRef]
3. Ezzat, M.A.; Ghany, M.A.A.E.; Almotairi, S.; Salem, M.A.-M. Horizontal Review on Video Surveillance for Smart Cities: Edge Devices, Applications, Datasets, and Future Trends. *Sensors* **2021**, *21*, 3222. [CrossRef] [PubMed]
4. Li, Y.; Qiao, Y.; Ruichek, Y. Multiframe-Based High Dynamic Range Monocular Vision System for Advanced Driver Assistance Systems. *IEEE Sens. J.* **2015**, *15*, 5433–5441. [CrossRef]
5. Barten, P.G.J. *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*; SPIE: Bellingham, WA, USA, 1999; Volume PM72.
6. Purohit, M.; Singh, M.; Kurmar, A.; Kaushik, B.K. Enhancing the Surveillance Detection Range of Image Sensors using HDR Techniques. *IEEE Sens. J.* **2021**, *21*, 19516–19528. [CrossRef]
7. Xu, Q.; Su, Z.; Zheng, Q.; Luo, M.; Dong, B. Secure Content Delivery with Edge Nodes to Save Caching Resources for Mobile Users in Green Cities. *IEEE Trans. Industr. Inform.* **2018**, *14*, 2550–2559. [CrossRef]
8. Yan, X.; Yin, P.; Tang, Y.; Feng, S. Multi-Keywords Fuzzy Search Encryption Supporting Dynamic Update in An Intelligent Edge Network. *Connect. Sci.* **2022**, *34*, 511–528. [CrossRef]
9. Ren, J.; Guo, H.; Xu, C.; Zhang, Y. Serving at the Edge: A Scalable IoT Architecture Based on Transparent Computing. *IEEE Netw.* **2017**, *31*, 96–105. [CrossRef]
10. Kalantari, N.K.; Ramamoorthi, R. Deep High Dynamic Range Imaging of Dynamic Scenes. *ACM Trans. Graph.* **2017**, *36*, 1–12. [CrossRef]
11. Wu, S.; Xu, J.; Tai, Y.-W.; Tang, C.-K. Deep High Dynamic Range Imaging with Large Foreground Motions. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 120–135.
12. Liu, Y.-L.; Lai, W.S.; Chen, Y.S.; Kao, Y.L.; Yang, M.H.; Chuang, Y.Y.; Huang, J.B. Single-Image HDR Reconstruction by Learning to Reverse the Camera Pipeline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1648–1657.
13. Eilertsen, G.; Kronander, J.; Denes, G.; Mantiuk, R.K.; Unger, J. HDR Image Reconstruction from A Single Exposure using Deep CNNs. *ACM Trans. Graph.* **2017**, *36*, 1–15. [CrossRef]
14. Banterle, F.; Ledda, P.; Debattista, K.; Chalmers, A. Inverse Tone Mapping. In Proceedings of the 4th International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia, Perth, Australia, 1–4 December 2006; pp. 349–356.
15. Landis, H. Production-Ready Global Illumination. In Proceedings of the International Conference on Computer Graphics and Interactive Techniques, San Antonio, TX, USA, 21–26 July 2002; pp. 93–95.
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Conference Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
17. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. SwinIR: Image Restoration using Swin Transformer. In Proceedings of the IEEE International Conference on Computer Vision Workshop, Montreal, BC, Canada, 11–17 October 2021; pp. 1833–1844.

18. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M. Restormer: Efficient Transformer for High-Resolution Image Restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 5718–5729.
19. Masia, B.; Agustin, S.; Fleming, R.W.; Sorkine, O.; Gutierrez, D. Evaluation of Reverse Tone Mapping through Varying Exposure Conditions. *ACM Trans. Graph.* **2009**, *28*, 1–8. [CrossRef]
20. Bist, C.; Cozot, R.; Madec, G.; Duclox, X. Tone Expansion using Lighting Style Aesthetics. *Comput. Graph.* **2017**, *62*, 77–86. [CrossRef]
21. Didyk, P.; Mantiuk, R.; Hein, M.; Seidel, H.P. Enhancement of Bright Video Features for HDR Displays. In *Computer Graphics Forum*; Blackwell Publishing Ltd.: Oxford, UK, 2008; pp. 1265–1274.
22. Rempel, A.G.; Trentacoste, M.; Seetzen, H.; Young, H.D.; Heidrich, W.; Whitehead, L.; Ward, G. LDR2HDR: On-The-Fly Reverse Tone Mapping of Legacy Video and Photographs. *ACM Trans. Graph.* **2007**, *26*, 39-es. [CrossRef]
23. Wu, G.; Song, R.; Zhang, M.; Li, X.; Rosin, P.L. LiTMNet: A Deep CNN for Efficient HDR Image Reconstruction from a Single LDR Image. *Pattern Recognit.* **2022**, *127*, 108620. [CrossRef]
24. Marnerides, D.; Bashford-Rogers, T.; Hatchett, J.; Debattista, K. ExpandNet: A Deep Convolutional Neural Network for High Dynamic Range Expansion from Low Dynamic Range Content. *Comput. Graph. Forum* **2018**, *37*, 37–49. [CrossRef]
25. Khan, Z.; Khanna, M.; Raman, S. FHDR: HDR Image Reconstruction from a Single LDR Image using Feedback Network. In Proceedings of the IEEE Global Conference on Signal and Information Processing, Ottawa, ON, Canada, 11–14 November 2019; pp. 1–5.
26. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3242–3250.
27. Fan, H.; Zhou, J. Stacked Latent Attention for Multimodal Reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1072–1080.
28. Yan, Q.; Gong, D.; Shi, Q.; van den Hengel, A.; Shen, C.; Reid, I.; Zhang, Y. Attention-Guided Network for Ghost-Free High Dynamic Range Imaging. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1751–1760.
29. Tel, S.; Wu, Z.; Zhang, Y.; Heyrman, B.; Demonceaux, C.; Timofte, R.; Ginhac, D. Alignment-Free HDR Deghosting with Semantics Consistent Transformer. In Proceedings of the IEEE International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 12836–12845.
30. Tao, H.; Duan, Q.; An, J. An Adaptive Interference Removal Framework for Video Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 5148–5159. [CrossRef]
31. Abdusalomov, A.B.; Mukhiddinov, M.; Whangbo, T.K. Brain Tumor Detection Based on Deep Learning Approaches and Magnetic Resonance Imaging. *Cancers* **2023**, *15*, 4172. [CrossRef] [PubMed]
32. Tao, H.; Duan, Q.; Lu, M.; Hu, Z. Learning Discriminative Feature Representation with Pixel-Level Supervision for Forest Smoke Recognition. *Pattern Recognit.* **2023**, *143*, 109761. [CrossRef]
33. Liu, S.; Zhang, X.; Sun, L.; Liang, Z.; Zeng, H.; Zhang, L. Joint HDR Denoising and Fusion: A Real-World Mobile HDR Image Dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 13966–13975.
34. Yan, Q.; Zhang, L.; Liu, Y.; Zhu, Y.; Sun, J.; Shi, Q.-F.; Zhang, Y. Deep HDR Imaging via a Non-Local Network. *IEEE Trans. Image Process.* **2020**, *29*, 4308–4322. [CrossRef]
35. 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 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
36. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jegou, H. Training Data-Efficient Image Transformers and Distillation through Attention. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; Volume 139, pp. 10347–10357.
37. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
38. Li, K.; Yu, R.; Wang, Z.; Yuan, L.; Song, G.; Chen, J. Locality Guidance for Improving Vision Transformers on Tiny Datasets. In Proceedings of the European Conference on Computer Vision, Tel-Aviv, Israel, 23–27 October 2022; pp. 110–127.
39. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. CvT: Introducing Convolutions to Vision Transformers. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 22–31.
40. Huang, T.; Li, S.; Jia, X.; Lu, H.; Liu, J. Neighbor2Neighbor: Self-Supervised Denoising from Single Noisy Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14781–14790.
41. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
42. Yin, P.; Yuan, R.; Cheng, Y.; Wu, Q. Deep Guidance Network for Biomedical Image Segmentation. *IEEE Access* **2020**, *8*, 116106–116116. [CrossRef]

43. Wang, L.; Yoon, K.-J. Deep Learning for HDR Imaging: State-of-the-Art and Future Trends. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 8874–8895. [CrossRef] [PubMed]
44. Zhang, J.; Wang, Y.; Tohidypour, H.; Pourazad, M.T.; Nasiopoulos, P. A Generative Adversarial Network Based Tone Mapping Operator for 4K HDR Images. In Proceedings of the IEEE International Conference on Computing, Networking and Communications, Honolulu, HI, USA, 20–22 February 2023; pp. 473–477.
45. Shen, H.; Zhang, H.; Shao, S.; Xin, J. Chromaticity-based separation of reflection components in a single image. *Pattern Recognit.* **2008**, *41*, 2461–2469. [CrossRef]
46. Johnson, J.; Alahi, A.; Li, F.F. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 694–711.
47. Funt, B.; Shi, L. The Effect of Exposure on MaxRGB Color Constancy. In Proceedings of the Human Vision and Electronic Imaging XV, San Jose, CA, USA, 18–21 January 2010.
48. Feng, X.; DiCarlo, J.; Catrysse, P.; Wandell, B. High dynamic range imaging of natural scenes. In Proceedings of the Color and Imaging Conference, Scottsdale, AZ, USA, 12–15 November 2002; pp. 337–342.
49. Reinhard, E.; Heidrich, W.; Debevec, P.; Pattanaik, S.; Ward, G.; Myszkowski, K. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*; Morgan Kaufmann: Burlington, MA, USA, 2010.
50. Fairchild, M. The HDR Photographic Survey. In Proceedings of the Color and Imaging Conference, Albuquerque, NM, USA, 5–9 November 2007; pp. 233–238.
51. Lee, B.; Sunwoo, M. HDR Image Reconstruction Using Segmented Image Learning. *IEEE Access* **2021**, *9*, 142729–142742. [CrossRef]
52. Zhou, C.; Smith, J.; Wang, Q.; Chen, L.; Wu, Z. Polarization Guided HDR Reconstruction via Pixel-Wise Depolarization. *IEEE Trans. Image Process.* **2023**, *32*, 1774–1787. [CrossRef]
53. Joffre, G.; Puech, W.; Comby, F.; Joffre, J. High Dynamic Range Images from Digital Cameras Raw Data. In Proceedings of the International Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 31 July–4 August 2005; p. 72-es.
54. Nemoto, H.; Korshunov, P.; Hanhart, P.; Ebrahimi, T. Visual Attention in LDR and HDR Images. In Proceedings of the International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Chandler, AZ, USA, 5–6 February 2015.
55. Kovaleski, R.; Oliveira, M. High-Quality Reverse Tone Mapping for a Wide Range of Exposures. In Proceedings of the Conference on Graphics, Patterns, and Images, Columbus, OH, USA, 26–30 August 2014; pp. 49–56.
56. Masia, B.; Serrano, A.; Gutierrez, D. Dynamic Range Expansion Based on Image Statistics. *Multimed. Tools Appl.* **2017**, *76*, 631–648. [CrossRef]
57. Wang, H.; Ye, M.; Zhu, X.; Li, S.; Zhu, C.; Li, X. KUNet: Imaging Knowledge-Inspired Single HDR Image Reconstruction. In Proceedings of the International Joint Conference on Artificial Intelligence, Vienna, Austria, 23–29 July 2022; pp. 1408–1414.
58. Hore, A.; Ziou, D.; Image Quality Metrics: PSNR, vs. SSIM. In Proceedings of the International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 2366–2369.
59. Wang, Z.; Simoncelli, P.; Bovik, C. Multiscale Structural Similarity for Image Quality Assessment. In Proceedings of the Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 9–12 November 2003; Volume 2, pp. 1398–1402.
60. Narwaria, M.; Mantiuk, R.; Da Silva, M.; Le Callet, P. HDR-VDP-2.2: A Calibrated Method for Objective Quality Prediction of High-Dynamic Range and Standard Images. *J. Electron. Imaging* **2015**, *24*, 010501. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.