Knowledge-Based Systems 263 (2023) 110278

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Structure-embedded ghosting artifact suppression network for high dynamic range image reconstruction

Lingfeng Tang [a,b], Huan Huang [a,b], Yafei Zhang [a,b,*], Guanqiu Qi [c,*], Zhengtao Yu [a,b]

[a] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, Yunnan, PR China
[b] Key Laboratory of Artificial Intelligence in Yunnan Province, Kunming University of Science and Technology, Kunming 650500, Yunnan, PR China
[c] State University of New York at Buffalo State, Buffalo, NY 14222, USA

ARTICLE INFO

Article history:
Received 24 July 2022
Received in revised form 4 January 2023
Accepted 5 January 2023
Available online 14 January 2023

Keywords:
High-dynamic-range imaging
Ghosting artifact suppression
Multi-head attention
Structure-embedded network

ABSTRACT

In high-dynamic-range (HDR) image reconstruction, the background offset among multiple multi-exposure low-dynamic-range (LDR) images, wide-range movement of targets, and missing edge structure information in the over-/under-exposure region cause both ghosting and blurring artifacts. This study proposed a structure-embedded ghosting artifact suppression network (SGARN) to achieve detailed preservation and ghosting artifact suppression to address this issue. According to the different image feature maps' correlation in channels, a channel and multi-head joint attention network (CMAN) was designed to highlight the features conducive to high-quality HDR image reconstruction. A dense multi-scale information transfer network (DMITN) was designed to integrate the characteristics of different combinations of convolution kernels with different receptive fields. In addition, a structure-embedded network was designed to predict the edge structure to be compensated from the reference image. The predicted edge was integrated into the reconstructed HDR image. Compared with state-of-the-art methods, the proposed method can achieve better visual performance and higher objective evaluation results on three public datasets. The source codes of the proposed method are available at https://github.com/lhf12278/SGARN.

1. Introduction

Compared to digital imaging, high-dynamic-range (HDR) imaging aims to reconstruct images with a higher dynamic range. Owing to its broad application prospects in games, virtual reality, medical imaging, and other fields, HDR imaging has attracted widespread research attention. In natural scenes, the dynamic range perceived by the human visual system is wide. The operating range of sensors directly constrains the performance of ordinary digital cameras. However, the dynamic imaging range of ordinary digital cameras is much smaller than the range the human eye perceives. High-end cameras are used as early-stage methods to capture HDR images. Although high-end cameras can capture HDR images, they are not only expensive but also not conducive to the development of HDR-related techniques. Therefore, acquiring high-quality HDR images has attracted considerable attention from researchers in recent years.

An HDR image can be reconstructed from a single low-dynamic-range (LDR) image as an effective method [1–7]. However, this type of method cannot integrate the complementary information of multiple LDR images captured in the same scene, resulting in poor HDR image quality. The information from multiple LDR images with different exposures captured in the same scene is quite different, and each image contains information from different dynamic range. Therefore, an HDR image can be obtained by fusing multiple LDR images with different exposure levels. Owing to various factors in real scenes, such as the offset of camera position and foreground target movement, the pixels of multiple multi-exposure LDR images captured in the same scene are not aligned in the spatial position. If these multi-exposure LDR images are directly fused using a static multi-exposure image fusion method, such as [8,9], noticeable ghosting artifacts are generated, thereby impairing the quality of the fused image.

Many effective methods have been proposed to generate high-quality HDR images by fusing unaligned multi-exposure images. Sen and Zimmer et al. [10,17] proposed to first align multiple LDR images and then eliminate the ghosting artifacts in the fusion process. This type of method relies excessively on the preprocessing step of image alignment. Therefore, the alignment performance is often poor when the related LDR images involve a complex scene and wide-range movement of objects, resulting in ghosting artifacts in the final fusion result. Heo and Jinno et al. [18–20] proposed first to detect motion areas, then predict areas with inconsistent information caused by the movement of targets, and

* Corresponding authors.
E-mail addresses: zyfeimail@kust.edu.cn (Y. Zhang), qig@buffalostate.edu (G. Qi).

https://doi.org/10.1016/j.knosys.2023.110278
0950-7051/

Fig. 1. Performance comparison of the reconstruction results of three LDR images involving the wide-range movement of the targets obtained by different methods. (a) three LDR images are to be fused with unaligned image contents. (b) the reconstruction result obtained by the proposed method. The images shown in the three rows below (a) and (b) compare three local areas extracted from three LDR images and the reconstruction results obtained by different methods. Three enlarged local areas (marked in red, blue, and green frames) extracted from LDR1, LDR2, LDR3, the reconstructed images obtained by (1) Sen [10], (2) HDRCNN [1], (3) SingleHDR [6], (4) Kalantari [11], (5) DeepHDR [12], (6) AHDRNet [13], (7) NHDRRNet [14], (8) HDRGAN [15], (9) HDRI [16] and (10) the proposed SGARN, and (11) the ground truth are shown from left to right.

finally discard them to solve the issue of information misalignment. This method does not fully use the information in non-reference images; therefore, the image reconstruction process cannot recover rich details.

In recent years, deep learning has achieved good performance in image fusion [21,22], image dehazing [23], person re-identification [24–27] and model parameter estimation [28,29]. Deep learning-based HDR reconstruction methods [12–16,30,31] can alleviate these issues to some extent. However, the lack of effective mining and utilization of input image features limits further quality improvement of the reconstruction results. Although attention mechanism-embedded deep learning methods [13] can effectively alleviate this issue, they do not highlight the role of the source image information in HDR image reconstruction at different feature levels. Therefore, their performance could not be improved further. As shown in Fig. 1, although some existing methods are effective, there is still room to improve ghosting artifact suppression and the maintenance and recovery of detailed information.

The proposed method solves the following three issues: (1) ghosting artifact suppression in reconstruction results caused by the misalignment of image pixels, (2) recovery of detailed information of over-/under-exposed regions, and (3) edge structure preservation of source images during the fusion process. From the perspective of image fusion, if the information conducive to improving the fusion quality is selected as much as possible for HDR image reconstruction, ghosting artifacts introduced by the misalignment of the source image information can be considerably suppressed. However, the loss of detailed information could be effectively alleviated. In addition, if edge details consistent with the ground truth can be predicted and generated based on the edge structure information of the reference image, the loss of over-/under-exposure image details can be effectively recovered. Therefore, this study proposes a structure-embedded ghosting artifact suppression network (SGARN) without registration to realize high-quality HDR image reconstruction by fusing multiple multi-exposure LDR images with shifted content.

Specifically, the correlation and importance of different image feature channels were fully considered. A channel and multi-head joint attention network (CMAN) is designed to achieve ghosting artifact suppression and highlight the role of information in improving the quality of reconstructed images. Unlike existing attention mechanisms, the proposed SGARN highlights the importance of source image information from multiple perspectives through a multi-head mechanism. Various characteristics can be extracted from various combinations of convolution kernels with different receptive fields. Therefore, a dense multi-scale information transfer network (DMITN) was designed to integrate various extracted characteristics to explore the input image information further. In addition, a structure-embedded mechanism is proposed to recover the lost over-/under-exposure image details and prevent the loss of detailed structural information from the source images. This mechanism can predict the detailed structure compensated by the reference image structure. The predicted result was integrated into the output of the DMITN to reconstruct an HDR image. The proposed network transfers source image information to the reconstructed image and recovers the lost over-/under-exposure image details, achieving ghosting artifact suppression (as shown in Fig. 1). The comparative experimental results demonstrate the proposed method's effectiveness and its superiority over state-of-the-art methods on three challenging public datasets.

This study has three main contributions as follows.

- A CMAN was proposed to highlight the importance of helpful information from different perspectives. This network fully considers the importance of feature channels and uses a multi-head mechanism to explore feature importance from different perspectives. Ghosting artifact suppression and both the enhancement and utilization of the source image information are effectively realized.

- A DMITN was designed to achieve both deep mining and utilization of useful information by employing the characteristics of different combinations of convolution kernels with different receptive fields. The role of helpful information is effectively highlighted to improve the reconstruction quality.
- A structure-embedded mechanism was proposed to prevent the loss of structural details associated with the expected HDR image. The structural details to be compensated can be predicted from the reference image structure and integrated into the output of the DMITN to reconstruct high-quality HDR images. This design effectively avoids the loss of details during the fusion process and recovers the lost over-/under-exposure image details.

The rest of this paper is organized as follows: Section 2 discusses the related work; Section 3 introduces the proposed method in detail; Section 4 analyzes the comparative experimental results; and Section 5 concludes this work.

2. Related work

According to the type of misalignment solution, HDR image reconstruction can be categorized into alignment-based methods [17,32–34], motion detection-based methods [18–20], and deep neural network-based methods [1,2,4,6,11–15,30].

2.1. Alignment-based methods

In image fusion, the fused images suffer from ghosting artifacts due to misalignment between the images to be fused. Therefore, it is necessary to perform a registration operation on the images to be fused before image fusion. Image registration is accurate guarantee of image fusion [35]. Commonly used image registration methods include optical flow-based methods [36,37], SIFT-based methods [38,39], local regions-based methods [40], etc. The multi-exposure image fusion method based on image registration is called alignment-based HDR image reconstruction. Alignment-based methods usually perform alignment processing on LDR images before performing multi-exposure image fusion. Ward et al. [32] proposed to align the shifted images by calculating the overall pixel offset through binary images. Tomaszewska et al. [33] proposed to use the SIFT algorithm to find the corresponding key points in consecutive frames to register different frames. Zimmer et al. [17] proposed an optical flow-based method to align images, but did not make use of HDR contents of misaligned regions. Hu et al. [34] proposed to realize alignment of target motion areas by dividing an image into blocks to find the connections among the information in the blocks. However, this type of method cannot usually achieve satisfactory alignment performance when any target/object has a wide-range movement in complex imaging scenes or LDR images, resulting in ghosting artifacts in fusion results.

2.2. Motion detection-based methods

Motion detection-based methods usually first register LDR images, then divide pixels into offset and unshifted ones, and finally eliminate offset pixels to prevent ghosting artifacts. Particularly, Jinno et al. [19] proposed to use Markov random field to estimate the displacement, occlusion, and saturation regions and exclude these regions in the final HDR image. Gallo et al. [41] proposed to first find the relationship between blocks in the logarithmic domain to identify motion areas, and then average the radiance estimation value to reconstruct an HDR image. Heo et al. [18] proposed to detect the regions involving ghosting artifacts caused by motion through a joint probability density function and use graph-cuts to optimize the energy function for eliminating ghosting artifacts. Raman et al. [42] used super-pixel grouping to detect scene changes and discarded blocks with inconsistent information to alleviate ghosting artifacts during HDR image reconstruction. Both template matching and hole filling were used by Zheng et al. [20] to detect and remove offset pixels, thereby suppressing ghosting artifact. The reconstructed images by such methods often have an LDR in motion areas, because they usually ignore motion areas rather than making full use of the information in motion areas.

2.3. Deep neural network-based methods

Deep neural network-based HDR imaging methods have two main types. (1) An HDR image is reconstructed from a single LDR image. (2) An HDR image is reconstructed by fusing multiple LDR images with different exposure levels. For the first type, Lee et al. [2] used adversarial learning to generate multiple multi-exposure LDR images from a single LDR image, and fused these generated LDR images to reconstruct an HDR image. Fotiadou et al. [4] applied sparse auto-encoders to model different exposure conditions from image block features, and then reconstructed an HDR image from a single LDR image by different exposure levels simulated. Eilertsen et al. [1] designed a deep autoencoder network to recover lost information in the saturated areas and enhance the details in the reconstructed image. Due to information loss in imaging process, Liu et al. [6] proposed an HDR-to-LDR image formation pipeline composed of dynamic range clipping, non-linear mapping from a camera response function, and quantization. Then, three convolutional neural networks (CNNs) were used to simulate inverse process of the above three steps to realize LDR-to-HDR transformation. Owing to relatively limited information in a single image, such methods usually cannot recover image details in the saturated areas.

For the second type, Kalantari et al. [11] first used an optical flow method to align input images, and then applied a deep neural network to reconstruct HDR image. Wu et al. [12] used an auto-encoder network to recover HDR image details during image translation. To further alleviate the negative influence of ghosting artifacts, Yan et al. [13] used a simple attention mechanism to guide network to suppress potential ghosting artifacts in non-reference frames. Yan et al. [14] applied feature space correlation to guide network to recover details in occluded areas. Prabhakar et al. [43] integrated both bilateral guided upsampling and motion segmentation masks to reconstruct high-resolution ghosting-artifact-free HDR images with limited computing resources. Although these methods can effectively improve the quality of the reconstructed images, there is still room to improvement in recovering detailed information lost in the over-/under-exposed areas or in suppressing ghosting artifacts introduced by wide range movement.

Different from both alignment-based and motion detection-based methods, the proposed method does not need to perform image alignment and motion detection in advance. It can automatically retrieve the information associated with the expected HDR image reconstruction result. Therefore, ghosting artifacts can be effectively suppressed. The proposed method is also different from existing deep neural network-based methods. It not only maximally explores the features from different levels and highlights the role of useful information in HDR images, but also predicts the structure consistent with the ground truth from the detailed edge information of a reference image. The recovery of lost details in over-/under-exposed areas can be realized.

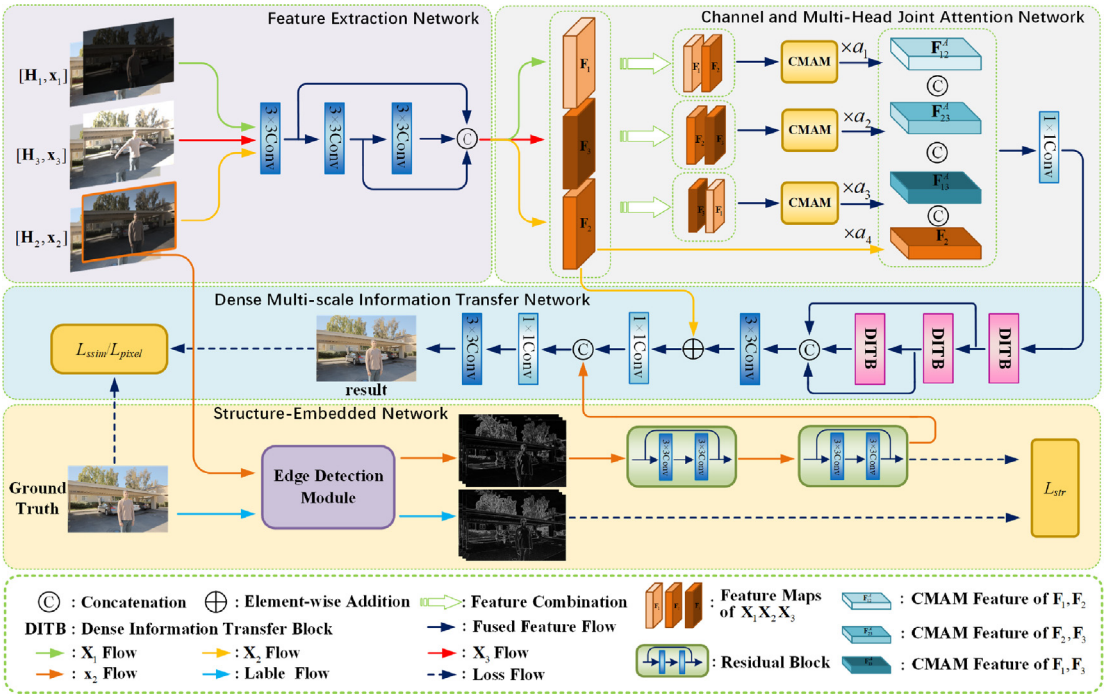

Fig. 2. Diagram of the proposed framework. The original input LDR images $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$ and the corresponding images $\{\boldsymbol{H}_1, \boldsymbol{H}_2, \boldsymbol{H}_3\}$ processed by gamma correction compose an input image group consisting of $\boldsymbol{X}_1 = [\boldsymbol{H}_1, \boldsymbol{x}_1]$, $\boldsymbol{X}_2 = [\boldsymbol{H}_2, \boldsymbol{x}_2]$, $\boldsymbol{X}_3 = [\boldsymbol{H}_3, \boldsymbol{x}_3]$. $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}$ is input into the feature extraction network to extract the initial feature maps. The obtained feature maps are concatenated on channels and sent to the CMAN. After processing in the CMAN, the relationship between feature channels of the same and different images can be explored and used, so the feature enhancement can be achieved. When the convolution kernels with different receptive fields are used in different combinations, they have different characteristics. In the DMITN, these characteristics are used to extract more information that is conducive to the high-quality HDR image reconstruction. The structure-embedded network (SEN) is applied to realize the recovery of the lost edge structure information.

3. Proposed method

3.1. Overview of the proposed framework

A set of multi-exposure LDR images $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$ captured in a dynamic scene is given. As the goal of HDR image reconstruction, an HDR image aligned with the selected reference image $\boldsymbol{x}_2$ is reconstructed according to the input non-reference images $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$, and the reconstructed image contains detailed information in $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$. According to the settings used in [11], LDR images $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$ to be fused are preprocessed by gamma correction to obtain the corresponding HDR images $\{\boldsymbol{H}_1, \boldsymbol{H}_2, \boldsymbol{H}_3\}$, and the HDR image $\boldsymbol{H}_i$ can be expressed as:

$$\boldsymbol{H}_i = \frac{\boldsymbol{x}_i^{\gamma}}{t_i}, i \in \{1, 2, 3\}, \tag{1}$$

where $\gamma > 1$ is gamma correction parameter. According to literature [44] and parameter analysis in Section 4.6, γ is set to 2.2. t_i is the exposure time of the LDR image $\boldsymbol{x}_i$. $\boldsymbol{x}_i$ and $\boldsymbol{H}_i$ are concatenated to obtain a tensor $\boldsymbol{X}_i = [\boldsymbol{x}_i, \boldsymbol{H}_i]$ with six channels. $\boldsymbol{X}_i = [\boldsymbol{x}_i, \boldsymbol{H}_i](i = 1, 2, 3)$ is used as the network input.

As shown in Fig. 2, the proposed method is mainly composed of four parts: a feature extraction network (FEN), a CMAN, a DMITN, and a structure-embedded network (SEN). The feature extraction network is mainly used to extract the initial features from the input images $\boldsymbol{X}_i, i \in \{1, 2, 3\}$. CMAN is mainly applied to highlight information conducive to improving image quality and reducing the negative influence of ghosting artifacts on the reconstruction result. DMITN fully uses the characteristics of different combinations of convolution kernels with different receptive fields to explore information from input images. SEN ensures detailed information in reference image is preserved during the reconstruction process and lost details in over-/under-exposure and motion regions are recovered.

3.2. Feature extraction network

As shown in Fig. 2, feature extraction is realized by three convolution layers. The output of each convolution layer has 16 channels. For an input $\boldsymbol{X}_i \in \mathbb{R}^{H \times W \times 6}$, the output of the nth convolution layer is expressed as:

$$\boldsymbol{F}_{c,i}^{n} = \text{conv}(\boldsymbol{X}_i, k = 3, n), \tag{2}$$

where k is convolution kernel size. The final output feature $\boldsymbol{F}_i \in \mathbb{R}^{H \times W \times 48}$ of the FEN is shown as follows:

$$\boldsymbol{F}_i = \text{concat}(\boldsymbol{F}_{c,i}^{1}, \boldsymbol{F}_{c,i}^{2}, \boldsymbol{F}_{c,i}^{3}). \tag{3}$$

where concat represents concatenation operation.

3.3. Channel and multi-head joint attention network

CMAN realizes enhancement and utilization of features by exploring the relationship between feature channels. This not only suppresses ghosting artifacts, but also is conducive to fully using the information contained in different LDR images to reconstruct high-quality HDR images. As shown in Fig. 3, CMAM consists of two branches, namely channel attention (CA) and multi-head attention (MA). CA is mainly used to highlight the importance of feature channels of a single image. To highlight salient information in feature maps, global average pooling (GAP) and global maximum pooling (GMP) are integrated. This study assumes the features of two different images input to CMAN are $\boldsymbol{F}_i$ and $\boldsymbol{F}_j$. In CA, the result obtained after concatenating $\boldsymbol{F}_i$ and $\boldsymbol{F}_j$ is first processed by 1×1 convolution, and then the corresponding GAP and GMP are applied to extract the global features $\boldsymbol{f}_{ij}^{g} \in \mathbb{R}^{1 \times C}$ and salient features $\boldsymbol{f}_{ij}^{h} \in \mathbb{R}^{1 \times C}$ on each channel respectively. The

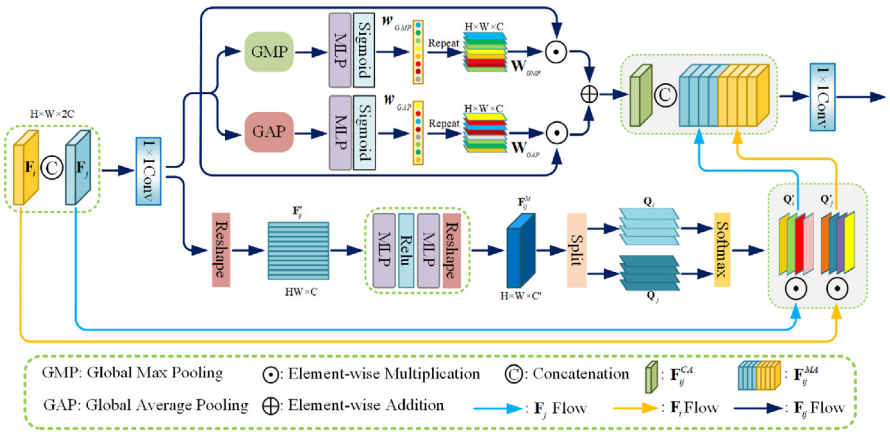

Fig. 3. Architecture of channel and multi-head joint attention mechanism (CMAM).

above process can be formulated as:

$$\begin{aligned} \boldsymbol{f}_{ij}^{g} &= \mathrm{GAP}(\mathrm{conv}(\mathrm{concat}(\boldsymbol{F}_i, \boldsymbol{F}_j), k=1)) \\ \boldsymbol{f}_{ij}^{h} &= \mathrm{GMP}(\mathrm{conv}(\mathrm{concat}(\boldsymbol{F}_i, \boldsymbol{F}_j), k=1))^{'} \end{aligned} \tag{4}$$

where $\boldsymbol{f}_{ij}^{g}$ and $\boldsymbol{f}_{ij}^{h}$ are input to the multi-layer perceptrons (MLP) φ_1 and φ_2 composed of two fully connected layers, respectively.

The weights $\boldsymbol{w}_{GAP} \in \mathbb{R}^{1\times1\times C}$ and $\boldsymbol{w}_{GMP} \in \mathbb{R}^{1\times1\times C}$ can be obtained by sigmoid activation function:

$$\begin{aligned} \boldsymbol{w}_{GAP} &= \mathrm{sigmoid}(\boldsymbol{\varphi}_1(\boldsymbol{f}_{ij}^{g})) \\ \boldsymbol{w}_{GMP} &= \mathrm{sigmoid}(\boldsymbol{\varphi}_2(\boldsymbol{f}_{ij}^{h}))^{'} \end{aligned} \tag{5}$$

According to $\boldsymbol{w}_{GAP}$ and $\boldsymbol{w}_{GMP}$, the feature $\boldsymbol{F}_{ij}^{CA} \in \mathbb{R}^{H\times W\times C}$ that integrates all the advantages of GAP and GMP is obtained:

$$\boldsymbol{F}_{ij}^{CA} = \mathrm{conv}(\boldsymbol{F}_{ij}, k=1) \odot \boldsymbol{W}_{GAP} + \mathrm{conv}(\boldsymbol{F}_{ij}, k=1) \odot \boldsymbol{W}_{GMP}, \tag{6}$$

where $\boldsymbol{F}_{ij} = \mathrm{concat}(\boldsymbol{F}_i, \boldsymbol{F}_j)$, $\odot$ represents element-wise multiplication. $\boldsymbol{W}_{GAP} \in \mathbb{R}^{H\times W\times C}$ and $\boldsymbol{W}_{GMP} \in \mathbb{R}^{H\times W\times C}$ are the tensors formed by copying each element of $\boldsymbol{w}_{GAP}$ and $\boldsymbol{w}_{GMP}$ into the matrix, respectively.

CA only focuses on the importance of different channel features of the fused feature maps, while ignoring the importance of different image feature maps at different levels. Therefore, the proposed method integrates MA into CA to explore the correlation between different images and highlight the role of useful information in image reconstruction.

To fully explore the channel connections between $\boldsymbol{F}_i$ and $\boldsymbol{F}_j$, MA first performs a reshape operation on the features processed after 1×1 convolution to obtain $\boldsymbol{F}_{ij}^{'}$ as follows:

$$\boldsymbol{F}_{ij}^{'} = \mathrm{reshape}(\mathrm{conv}(\boldsymbol{F}_{ij}, k=1)). \tag{7}$$

$\boldsymbol{F}_{ij}^{'} \in \mathbb{R}^{HW\times C}$ is sent to the multi-layer perceptions φ_3 and φ_4 to explore the relationship between different channel features. φ_3 and φ_4 are composed of two fully connected layers, respectively. The output of this process is expressed as:

$$\tilde{\boldsymbol{F}}_{ij} = \varphi_4(\mathrm{ReLU}(\varphi_3(\boldsymbol{F}_{ij}^{'}))). \tag{8}$$

To highlight the features conducive to the reconstruction result, $\tilde{\boldsymbol{F}}_{ij}$ is first reshaped to obtain $\boldsymbol{F}_{ij}^{M} \in \mathbb{R}^{H\times W\times C^{'}}$. Then, $\boldsymbol{F}_{ij}^{M}$ is divided into two sets of feature maps $\boldsymbol{Q}_i \in \mathbb{R}^{H\times W\times C^{'}/2}$ and $\boldsymbol{Q}_j \in \mathbb{R}^{H\times W\times C^{'}/2}$ at channel level. Next, softmax operation is performed on them respectively to obtain the multi-head self-attention weight maps $\boldsymbol{Q}_i^{'}$ and $\boldsymbol{Q}_j^{'}$:

$$\{\boldsymbol{Q}_i^{'}, \boldsymbol{Q}_j^{'}\} = \mathrm{softmax}(\mathrm{split}(\boldsymbol{F}_{ij}^{M})), \tag{9}$$

where $\boldsymbol{F}_{ij}^{M} = \mathrm{reshape}(\tilde{\boldsymbol{F}}_{ij}) \in \mathbb{R}^{H\times W\times C^{'}}$, $C^{'}$ is the number of heads, and split is the even division operation of feature maps by channels.

Existing multi-head self-attention methods directly multiply weight maps and input feature maps. However, the proposed method separates the multi-head attention weight maps. The separated weight maps are multiplied with the corresponding elements of each channel of the input features $\boldsymbol{F}_i$ and $\boldsymbol{F}_j$ to obtain the feature map set $\{\boldsymbol{F}_i^1, \boldsymbol{F}_i^2, \ldots, \boldsymbol{F}_i^{C^{'}/2}\}$:

$$\begin{aligned} \boldsymbol{F}_i^m &= \boldsymbol{Q}_i^{'}\{m\} \odot \boldsymbol{F}_i \\ \boldsymbol{F}_j^m &= \boldsymbol{Q}_j^{'}\{m\} \odot \boldsymbol{F}_j^{'} \end{aligned} \tag{10}$$

where $m = 1, 2, \ldots, C^{'}/2$, $\boldsymbol{Q}_i^{'}\{m\} \in \mathbb{R}^{H\times W}$ and $\boldsymbol{Q}_j^{'}\{m\} \in \mathbb{R}^{H\times W}$ are the weight maps of the mth head of $\boldsymbol{F}_i$ and $\boldsymbol{F}_j$ respectively. $C^{'}$ feature maps are concatenated to obtain the feature map $\boldsymbol{F}_{ij}^{MA}$ of the multi-head self-attention branch as follows:

$$\boldsymbol{F}_{ij}^{MA} = \mathrm{concat}(\boldsymbol{F}_i^1, \boldsymbol{F}_i^2, \ldots, \boldsymbol{F}_i^{C^{'}/2}, \boldsymbol{F}_j^1, \boldsymbol{F}_j^2, \ldots, \boldsymbol{F}_j^{C^{'}/2}), \tag{11}$$

where $\boldsymbol{F}_{ij}^{MA} \in \mathbb{R}^{H\times W\times(C\times C^{'})}$. In this way, the relationship between the features of different channels in different images can be fully explored to reconstruct HDR images, which is conducive to highlighting the information associated with the expected HDR image and suppressing ghosting artifacts.

To integrate the advantages of CA and MA, the output feature $\boldsymbol{F}_{ij}^{CA}$ of CA and the output feature $\boldsymbol{F}_{ij}^{MA}$ of MA are first concatenated, and then fused through 1×1 convolution. Finally, the feature $\boldsymbol{F}_{ij}^{A} \in \mathbb{R}^{H\times W\times K}$ of CMAN-guided feature is obtained:

$$\boldsymbol{F}_{ij}^{A} = \mathrm{conv}(\mathrm{concat}(\boldsymbol{F}_{ij}^{CA}, \boldsymbol{F}_{ij}^{MA}), k=1). \tag{12}$$

Assuming that there are three original images to be fused, the feature maps $\boldsymbol{F}_{12}^{A}$, $\boldsymbol{F}_{23}^{A}$, and $\boldsymbol{F}_{13}^{A}$, and the feature map $\boldsymbol{F}_2$ of the reference image are obtained by CMAN. $\boldsymbol{F}_{12}^{A}$, $\boldsymbol{F}_{23}^{A}$, and $\boldsymbol{F}_{13}^{A}$ contain the relationship between any two images. Considering the difference in information carried by $\boldsymbol{F}_{12}^{A}$, $\boldsymbol{F}_{23}^{A}$, $\boldsymbol{F}_{13}^{A}$, and $\boldsymbol{F}_2$, the learnable weights α_1, α_2, α_3, and α_4 are introduced to adjust the role of $\boldsymbol{F}_{12}^{A}$, $\boldsymbol{F}_{23}^{A}$, $\boldsymbol{F}_{13}^{A}$, and $\boldsymbol{F}_2$ in HDR image reconstruction. These feature maps is integrated as follows:

$$\boldsymbol{F}_{sum} = \mathrm{concat}(\alpha_1\boldsymbol{F}_{12}^{A}, \alpha_2\boldsymbol{F}_{23}^{A}, \alpha_3\boldsymbol{F}_{13}^{A}, \alpha_4\boldsymbol{F}_2). \tag{13}$$

3.4. Dense multi-scale information transfer network

To ensure the high quality of the reconstructed HDR image, an effective feature extraction method is first used to maximally extract the useful information carried by source images, and then the extracted information is applied to HDR image reconstruction. For convolution-based feature extraction, the extracted features

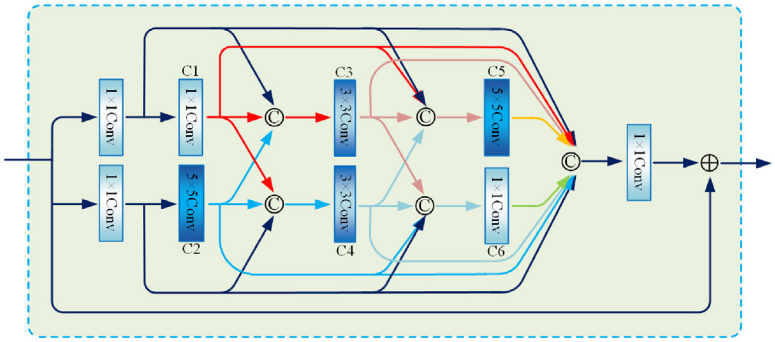

Fig. 4. Structure of dense information transfer block (DITB) in DMITN.

are different when different combinations of convolution kernels with different receptive fields are used. For example, after applying the 3 × 3 convolution layer to the output of the 5 × 5 convolution layer, the extracted features show the characterization of the image features extracted from the large receptive field (5 × 5) in the small receptive field (3 × 3). In contrast, after applying the 5 × 5 convolution layer to the output features of the 3 × 3 convolution layer, the extracted features show the characterization of the image features extracted from the small receptive field (3 × 3) in the large receptive field (5 × 5).

These two feature extraction methods can describe image features from different aspects. The features extracted by them have a certain complementarity. Therefore, a dense information transfer block (DITB) is proposed as shown in Fig. 4, which consists of two branches, upper and lower. The upper branch is composed of convolutional layers C1, C3 and C5, and the lower branch is composed of convolutional layers C2, C4 and C6. The convolution kernels are set to 1 × 1, 3 × 3, and 5 × 5 for C1, C3, and C5, respectively, and 5 × 5, 3 × 3, and 1 × 1 for C2, C4, and C6, respectively. In addition, dense skip connections are added to the DITB, which helps prevent the loss of shallow features. Specifically, the convolution layers are divided into three groups, C1 and C2, C3 and C4, and C5 and C6. After the input feature $\tilde{\boldsymbol{F}}_{sum1}$ and $\tilde{\boldsymbol{F}}_{sum2}$ are processed by 1 × 1 and 5 × 5 convolutions respectively, the corresponding features $\boldsymbol{F}_{c1}$ and $\boldsymbol{F}_{c2}$ are obtained:

$$\boldsymbol{F}_{c1} = \mathrm{conv}(\tilde{\boldsymbol{F}}_{sum1}, k = 1), \boldsymbol{F}_{c2} = \mathrm{conv}(\tilde{\boldsymbol{F}}_{sum2}, k = 5), \tag{14}$$

where $\tilde{\boldsymbol{F}}_{sum1} = \mathrm{conv}_1(\boldsymbol{F}_{sum}, k = 1)$, $\tilde{\boldsymbol{F}}_{sum2} = \mathrm{conv}_2(\boldsymbol{F}_{sum}, k = 1)$. conv_1 and conv_2 denote two different 1 × 1 convolutions.

In the upper and lower branches, the results obtained by concatenation operation can be expressed as:

$$\begin{aligned} \boldsymbol{F}_{up1} &= \mathrm{concat}(\boldsymbol{F}_{c1}, \boldsymbol{F}_{c2}, \tilde{\boldsymbol{F}}_{sum1}) \\ \boldsymbol{F}_{dow1} &= \mathrm{concat}(\boldsymbol{F}_{c2}, \boldsymbol{F}_{c1}, \tilde{\boldsymbol{F}}_{sum2}) \end{aligned}, \tag{15}$$

The features obtained by the first concatenation operation are further extracted using two 3 × 3 convolution layers, and then the second concatenation operation is performed:

$$\begin{aligned} \boldsymbol{F}_{up2} &= \mathrm{concat}(\boldsymbol{F}_{c3}, \boldsymbol{F}_{c1}, \tilde{\boldsymbol{F}}_{sum1}, \boldsymbol{F}_{c4}) \\ \boldsymbol{F}_{dow2} &= \mathrm{concat}(\boldsymbol{F}_{c3}, \boldsymbol{F}_{c2}, \tilde{\boldsymbol{F}}_{sum2}, \boldsymbol{F}_{c4}) \end{aligned}, \tag{16}$$

where $\boldsymbol{F}_{c3} = \mathrm{conv}(\boldsymbol{F}_{up1}, k = 3)$, $\boldsymbol{F}_{c4} = \mathrm{conv}(\boldsymbol{F}_{dow1}, k = 3)$.

After dense information transfer, the concatenated feature is obtained:

$$\boldsymbol{F}_{con} = \mathrm{concat}(\boldsymbol{F}_{c5}, \boldsymbol{F}_{c3}, \boldsymbol{F}_{c1}, \tilde{\boldsymbol{F}}_{sum1}, \tilde{\boldsymbol{F}}_{sum2}, \boldsymbol{F}_{c6}, \boldsymbol{F}_{c4}, \boldsymbol{F}_{c2}), \tag{17}$$

where $\boldsymbol{F}_{c5} = \mathrm{conv}(\boldsymbol{F}_{up2}, k = 5)$, $\boldsymbol{F}_{c6} = \mathrm{conv}(\boldsymbol{F}_{dow2}, k = 1)$. $\boldsymbol{F}_{con}$ is processed by 1 × 1 convolution to achieve channel fusion, and the input feature $\boldsymbol{F}_{sum}$ is added. The feature $\tilde{\boldsymbol{F}}_{merg}$ is obtained:

$$\tilde{\boldsymbol{F}}_{merg} = \mathrm{conv}(\boldsymbol{F}_{con}, k = 1) + \boldsymbol{F}_{sum}, \tag{18}$$

where $\tilde{\boldsymbol{F}}_{merg}$ is the feature obtained by a DITB. The proposed method uses a total of three DITBs to extract useful information for HDR image reconstruction.

In the DMITN, the output features of three DITBs are concatenated. The concatenated features are first processed by 3 × 3 convolution, and the feature $\boldsymbol{F}_2$ of the reference image is added. Then, the processed features and $\boldsymbol{F}_2$ are fused by 1 × 1 convolution. The fused feature maps are concatenated with the compensation features predicted by the structure-embedded network. Finally, 1 × 1 convolution is used to fuse the concatenated features on channels, and 3 × 3 convolution is applied for reconstructing HDR image.

3.5. Structure-embedded network

In LDR images, some detail is lost in the target motion areas and over-/under-exposure areas. In HDR image reconstruction, it is difficult to directly recover lost details in these areas. This study attempts to use the edge structure in the reference image to predict the detailed structure of the expected HDR image. Therefore, SEN is proposed to compensate the detailed structure information in the reference image to the reconstructed HDR image.

In SEN, sobel operator is first used to extract edge information of the three channels R, G, and B of the input image, and then the extracted edge information is concatenated to obtain a three-channel structure edge map. The concatenated edge image is mapped by residual blocks to obtain the edge features $\boldsymbol{F}_{edge}$ that need to be integrated. The above process can be expressed as:

$$\boldsymbol{F}_{edge} = \mathrm{RB}(\mathrm{concat}(\mathrm{soble}(\boldsymbol{x}_2^R), \mathrm{soble}(\boldsymbol{x}_2^G), \mathrm{soble}(\boldsymbol{x}_2^B))), \tag{19}$$

where $\boldsymbol{x}_2^R$, $\boldsymbol{x}_2^G$, $\boldsymbol{x}_2^B$ represent the R, G, and B channels of the reference image $\boldsymbol{x}_2$ respectively, and RB represents residual blocks.

Unlike existing HDR reconstruction methods that directly recover lost information from over-/under-exposure and target motion areas, the proposed method predicts the detailed structure that needs to be recovered based on the reference image $\boldsymbol{x}_2$. The proposed method not only makes full use of the edge structure information in the reference image, but also effectively avoids the problem of recovering lost details directly from over-/under-exposure and target motion areas. Therefore, the proposed method has a stronger feature recovery ability compared to existing methods.

3.6. Loss function

According to the settings in literature [11], both ground truth HDR image $\hat{\boldsymbol{H}}$ and network prediction image $\boldsymbol{H}$ are processed by the tone mapping function before calculating the loss. The differentiable μ-law function for tone mapping is defined as:

$$T(\boldsymbol{H}) = \frac{\log(1 + \mu \boldsymbol{H})}{\log(1 + \mu)}, \tag{20}$$

where μ as a parameter determines the degree of compression, which is set to 5000 according to [11]. $T(\boldsymbol{H})$ is the tone mapped image of the HDR image $\boldsymbol{H}$. In the proposed method, l_1-loss is used to ensure that the reconstructed HDR image $\boldsymbol{H}$ is consistent with the ground truth $\hat{\boldsymbol{H}}$:

$$L_{pixel} = \|T(\boldsymbol{H}) - T(\hat{\boldsymbol{H}})\|_1, \tag{21}$$

where $T(\hat{\boldsymbol{H}})$ represents the ground truth after tone mapping.

In the SEN, l_1-loss is applied to the output of SEN and edge structure of the ground truth. Therefore, the network can accurately predict the edge details that need to be compensated for the reconstructed HDR image:

$$L_{str} = \|\boldsymbol{F}_{label-str} - \boldsymbol{F}_{str}\|_1, \tag{22}$$

where $\boldsymbol{F}_{label-str}$ is the edge structure of $\hat{\boldsymbol{H}}$ detected by soble operator, and $\boldsymbol{F}_{str}$ is the output of SEN.

The structural similarity (SSIM) [45] is used to measure the structural similarity between two images. When the value of SSIM increases, the structure of the two images becomes more similar. The loss function L_{ssim} is used to ensure that the structural similarity between the reconstructed image and ground truth is high.

$$\begin{aligned} L_{ssim} &= 1 - \mathrm{SSIM}(T(\boldsymbol{H}), T(\hat{\boldsymbol{H}})) \\ &= 1 - \frac{(2\mu_{T(\boldsymbol{H})}\mu_{T(\hat{\boldsymbol{H}})} + \tau_1)(2\sigma_{T(\boldsymbol{H})T(\hat{\boldsymbol{H}})} + \tau_2)}{(\mu^2_{T(\boldsymbol{H})} + \mu^2_{T(\hat{\boldsymbol{H}})} + \tau_1)(\sigma^2_{T(\boldsymbol{H})} + \sigma^2_{T(\hat{\boldsymbol{H}})} + \tau_2)}, \end{aligned} \tag{23}$$

where $\mu_{T(\boldsymbol{H})}$ and $\mu_{T(\hat{\boldsymbol{H}})}$ represent the mean values of $T(\boldsymbol{H})$ and $T(\hat{\boldsymbol{H}})$ respectively. $\sigma_{T(\boldsymbol{H})T(\hat{\boldsymbol{H}})}$ represents the covariance of $T(\boldsymbol{H})$ and $T(\hat{\boldsymbol{H}})$. $\sigma^2_{T(\boldsymbol{H})}$ and $\sigma^2_{T(\hat{\boldsymbol{H}})}$ represent the variance of $T(\boldsymbol{H})$ and $T(\hat{\boldsymbol{H}})$ respectively. τ_1 and τ_2 are constants. The total loss function can be expressed as:

$$L = L_{pixel} + \alpha L_{str} + \beta L_{ssim}. \tag{24}$$

where α and β are hyperparameters.

4. Experiments

4.1. Dataset

Training dataset: Kalantari's dataset [11] is commonly used in HDR image reconstruction, including training and test sets. The training set of Kalantari's dataset is used to train the model in this study. Specifically, the training dataset contains 74 image sets taken in different scenes, and all images is resized to 1500×1000. Each image set contains three LDR images with different exposure levels and one HDR image captured in the same scene. In each image set, the medium exposure image among three LDR images is selected as the reference image, and the HDR image is used as the ground truth. The reference image and ground truth contain the same scene content, and the two images are strictly aligned. To increase the number of the training samples and avoid model overfitting, the training images are randomly cropped, horizontally and vertically flipped using the method in [12] to obtain 512×512 image blocks as the training samples of the proposed model.

Testing dataset: This study uses Kalantari's [11], Sen's [10] and Tursun's [46] datasets to verify the effectiveness of the proposed method. Kalantari's dataset [11] includes 15 testing sample sets, and each set contains three LDR images to be fused and one ground truth HDR image. Sen's [10] and Tursun's [46] datasets without ground truth are used as the testing sets to verify the generalization ability of the proposed method. For these two datasets, the comparative experimental results show the visual effect of the reconstructed HDR images obtained by different methods, and blind image quality assessment is performed on the reconstructed images.

4.2. Implementation details

In the training phase, Adam optimizer [47] is used, and batch size is set to 3. The learning rate is adjusted by warm-up to achieve the effective network training. First, the initial learning rate is set to 10^{-3} and used until the 1500th epoch. Then, the learning rate decays to 10^{-4}, and to 10^{-5} at the 2500th epoch. Finally, 10^{-5} as the learning rate is used until the end of training. The proposed model is trained for a total of 3100 epochs. PyTorch is used as the implementation platform of the proposed framework. The proposed entire network model is trained on a PC equipped with an Intel i10 CPU, 36 GB memory, and an Nvidia GeForce RTX 3090 GPU.

4.3. Evaluation metrics

PSNR-L, SSIM-L, PSNR-μ, SSIM-μ, and HDR-VDP-2 [48] are used to evaluate the reconstruction results obtained by different methods on Kalantari's testing dataset. In addition, UDQM [46] and BTMQI [49] are two metrics of blind image quality assessment to measure the performance of different methods on Sen's and Tursun's datasets. PSNR-L and SSIM-L are used to measure the similarity between the reconstructed HDR image and ground truth in the linear domain. PSNR-μ and SSIM-μ are used to evaluate the quality of the reconstructed HDR image in the μ-law domain. HDR-VDP-2 is used to measure the visibility and quality of the reconstructed HDR image under different brightness conditions. UDQM is used to comprehensively evaluate the ghosting artifact suppression of the reconstructed HDR image from four aspects Q_B (blending metric), Q_G (gradient inconsistency metric), Q_V (visual difference metric), and Q_D (dynamic range metric). BTMQI is used to evaluate the overall image quality by measuring the information entropy, naturalness, and structure information of tonemapped HDR images.

4.4. Comparison with state-of-the-art methods

The proposed method is compared with the state-of-the-art methods on Kalantari's, Sen's, and Tursun's testing datasets to verify its effectiveness. The comparative methods can be divided into three categories: (1) patch-based method, Sen [10], (2) single-frame reconstruction-based methods, HDRCNN [1] and SingleHDR [6], (3) deep neural network-based methods, Kalantari [11], DeepHDR [12], AHDRNet [13], NHDRRNet [14], HDR-GAN [15] and HDRI [16]. Specifically, Kalantari's method requires optical flow to preprocess image alignment before images are fed to the network, and DeepHDR needs to use homography transformation to align the background of the input images. AHDRNet, NHDRRNet, HDRGAN, HDRI and the proposed method do not require any image preprocessing.

Visual evaluation on dataset with ground truth: Figs. 5 and 6 show the reconstruction results obtained by different methods on two testing sample sets of Kalantari's dataset. All HDR images displayed in this section are tonemapped using Photomatix [12]. Figs. 5(a)–(c) and 6(a)–(c) are LDR images. Owing to the movement of the foreground targets and the over-/under-exposure of local areas in the images to be fused, some local areas in the images lose partial details. This poses a great challenge for high-quality HDR image reconstruction. To recover lost details based on the reference image (shown in Figs. 5(b) and 6(b)), the information in LDR images must be integrated into the reconstructed HDR image. In addition, due to the movement of targets, ghosting artifacts appear in the fusion result. Therefore, ghosting artifact suppression and image edge detail restoration and maintenance are two key issues that should be solved in HDR image reconstruction. Figs. 5(d)–(m) and 6(d)–(m) compare the

Fig. 5. Reconstructed HDR images obtained by different methods on the parking lot scene-1 image in the testing dataset of Kalantari dataset. (a)–(c) are three LDR images with different exposures and (b) is selected as the reference image. (d)–(m) are HDR images reconstructed by different methods. (n) is the ground truth.

visual effects of tonemapped HDR images obtained by different methods. Partial local areas of these HDR images are enlarged for further comparison.

As shown in Fig. 5(d), although Sen's method can reconstruct a HDR image, it cannot ideally recover the lost details. Additionally, when this method fuses the images shown in Fig. 6(a)–(c), it not only causes unsatisfactory detail recovery (shown in the enlarged background area marked in the red frame of Fig. 6(d)), but also introduces distortion (shown in the enlarged area marked in the blue frame of Fig. 6(d)). As a main reason, this block-based method mismatches some blocks from the saturated areas of different images. Although HDRCNN and SingleHDR as two single-frame reconstruction-based methods can avoid the introduction of ghosting artifacts and distortion, they cannot obtain necessary information from non-reference images to reconstruct edge details. Therefore, the reconstructed images have blurred artifacts. As shown in the areas marked in blue and green frames of Figs. 5(g) and 6(g), the reconstructed image obtained by Kalantari's optical flow-based method has color distortion and ghosting artifacts, which are caused by the deviation of optical flow alignment.

Although DeepHDR and NHDRRNet achieve better visual effects, slight ghosting artifacts appear on human face, which are marked in the green frames of Fig. 5(h) and (j). Additionally, the enlarged areas marked in red frames of Fig. 6(h) and (j) are over-bright, and the partial details of the objects appearing at the local image edges are weakened. Although AHDRNet introduces an attention mechanism to alleviate ghosting artifacts to a certain extent, it fails to highlight the role of key information in non-reference images. Therefore, slight ghosting and blurring artifacts appear in the enlarged areas marked in the red and green frames of Fig. 5(i). Additionally, as shown in the enlarged areas marked in the red and green frames of Fig. 6(i), not only oversaturation is introduced, but also some image details are lost. Owing to the unstable training of Generative Adversarial Network (GAN), the results of the HDRGAN method suffer from slight ghosting artifacts (shown in the region marked in the blue frame in Fig. 5(k) and the region marked in the green frame in Fig. 6(k)). According to the region marked in the blue frame in Fig. 5(l), ghosting artifacts appear in the experimental results of the HDRI method due to the serious oversaturation area in the scene. Compared with the above existing methods, the proposed method highlights the useful information and predicts the detailed structure information of the expected HDR image based on the edge structure of the reference images. Therefore, the proposed method can effectively suppress ghosting artifacts and recover image details.

Visual evaluation on dataset without ground truth: This study compares the performance of different methods on Sen's and Tursun's datasets to verify the generalization ability of the proposed model. The reconstruction results of two scenes of each dataset are demonstrated. Fig. 7 show the four sets of LDR images to be fused from Sen's and Tursun's datasets. Fig. 8 show the

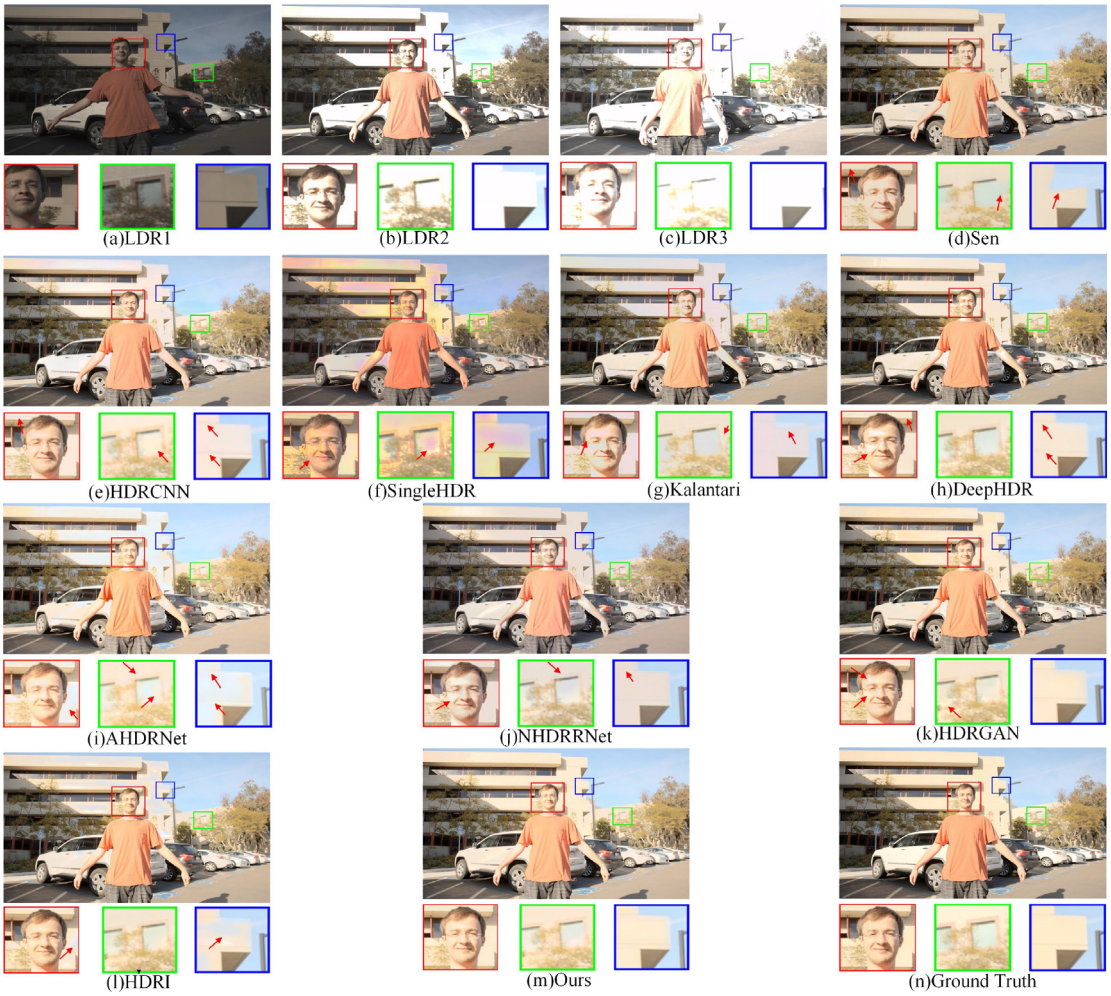

Fig. 6. Reconstructed HDR images obtained by different methods on the parking lot scene-2 image in the testing dataset of Kalantari dataset. (a)–(c) are three LDR images with different exposures and (b) is selected as the reference image. (d)–(m) are HDR images reconstructed by different methods. (n) is the ground truth.

tonemapped HDR images obtained by different methods on the two scenes of Sen's dataset shown in Fig. 7. As shown in the enlarged area marked in the red frame of Fig. 8(a), owing to the movement of the hand in the non-reference images, Sen's method introduces ghosting artifacts. Additionally, this method not only fails to effectively recover the handwriting shown in the enlarged areas marked in the red frame of Fig. 8(k), but also introduces slight noise in the lower jaw. The single-frame reconstruction-based methods HDRCNN and SingleHDR cannot integrate the information of all LDR images, resulting in poor performance on the recovery of image details and even partial over-exposure (as shown in Fig. 8(b) and (c)).

Owing to the simple network structure, Kalantari's method not only introduces ghosting artifacts shown in the enlarged areas of Fig. 8(d), but also fails to effectively recover the details. As shown in the enlarged area marked in the green frame of Fig. 8(n), this method also introduces noise into the reconstruction result. In addition, as shown in the enlarged areas marked in the red frames of Fig. 8(e)–(g) and (o)–(q), DeepHDR, AHDRNet, and NHDRRNet fail to recover lost details. DeepHDR and NHDRRNet also introduce slight grid effect in the fusion results (shown in the enlarged area marked in the green frames of Fig. 8(o) and (q)). DeepHDR and NHDRRNet are developed based on the U-net structure, deconvolution operation tends to introduce grid effect. The attention-guided method AHDRNet does not make full use of the information in the non-reference images, resulting in unsatisfactory details of the reconstruction results. Although HDRGAN achieves better visual effect, it introduces slight noise as shown in the marked region in Fig. 8(h). According to the region marked in red in Fig. 8(s), noise appears in the experimental results of HDRI, when dealing with scenes with a large dynamic range. In contrast, the proposed method fully explores the useful information of all LDR images, and makes full use of the edge details of the reference images, so high-quality HDR images can be reconstructed.

Fig. 9 shows the reconstruction results obtained by different methods on the two sets of LDR images shown in the last two rows of Fig. 7. Sen's method introduces ghosting artifacts shown in the enlarged area marked in green frame of Fig. 9(a) and (k), which affect the visual effect of the reconstruction results. HDRCNN and SingleHDR cause the color distortion of the reconstruction results. As shown in Fig. 9(d) and (n), the reconstruction results obtained by Kalantari's method have severe deformation and ghosting artifacts (shown in the enlarged area marked in green frame of Fig. 9(d) and (n)). As a main reason, Kalantari's method cannot solve the misalignment caused by the optical flow method. As shown in Fig. 9(e), (f), (o), and (p), the results obtained by DeepHDR and AHDRNet not only cause the distortion of edge details but also introduce ghosting artifacts. Mesh effect is introduced to the reconstruction results obtained by DeepHDR and NHDRRNet, and ghosting artifacts as shown in Fig. 9(o) and (q) are not effectively suppressed. The region marked in green in Fig. 9(h) is slightly blurred. This is because HDRGAN uses a cascaded structure to design the generator, which may cause the

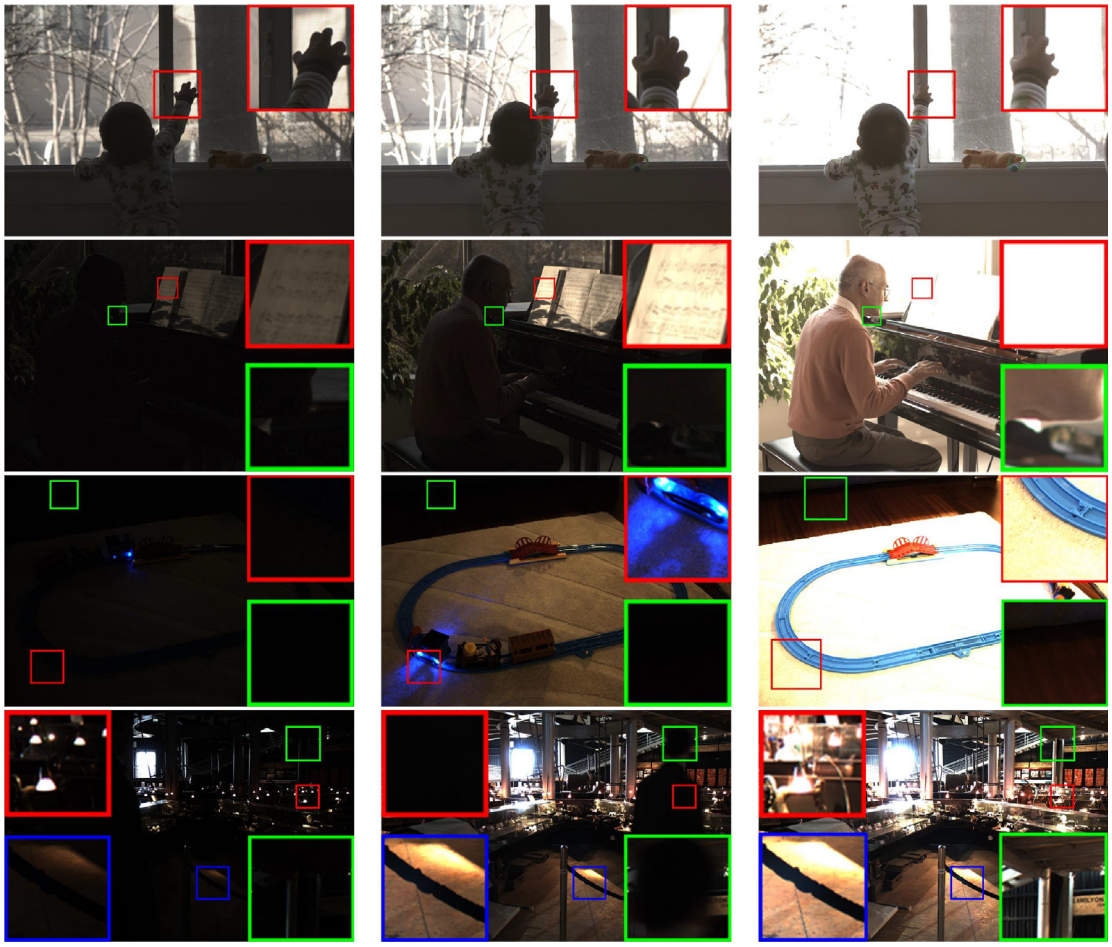

Fig. 7. Four sets of LDR images from Sen's and Tursun's datasets. LDR1, LDR2 and LDR3 images are shown from left to right. The first and second row are from Sen's dataset, and the third and fourth row are from Tursun's dataset.

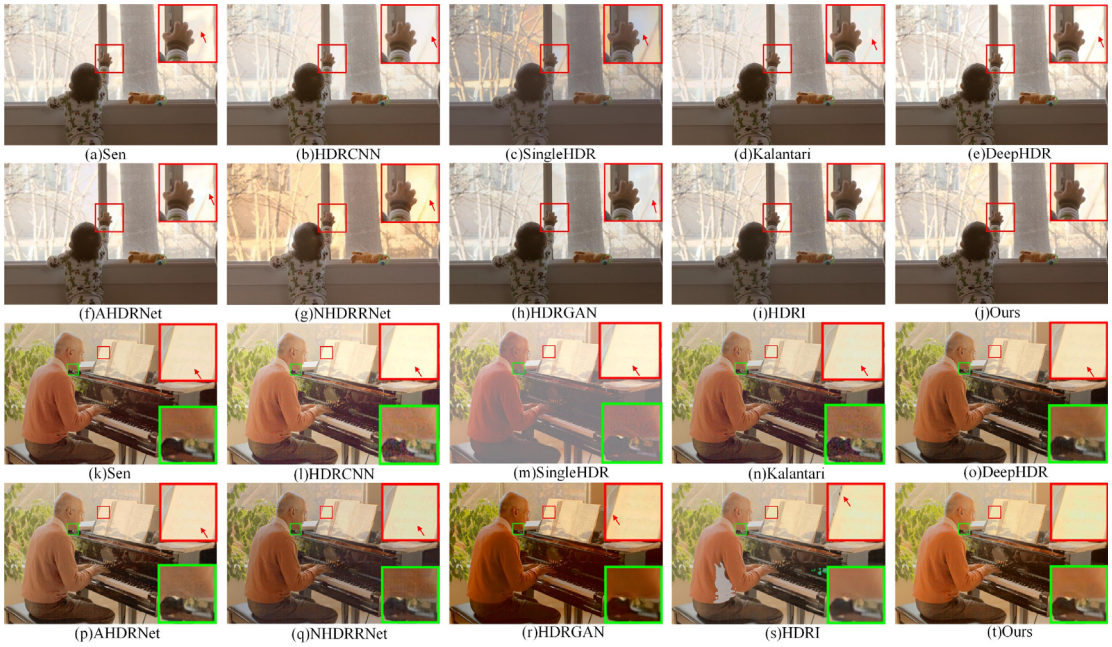

Fig. 8. Comparison of the reconstruction results obtained by different methods on the two sets of LDR images from Sen's dataset.

loss of some details in the reconstruction process. As shown in the regions marked in red and green in Fig. 9(i), HDRI eliminates the ghosting artifacts caused by the movement of the car but cannot recover the details of the over/under-saturated areas, resulting in

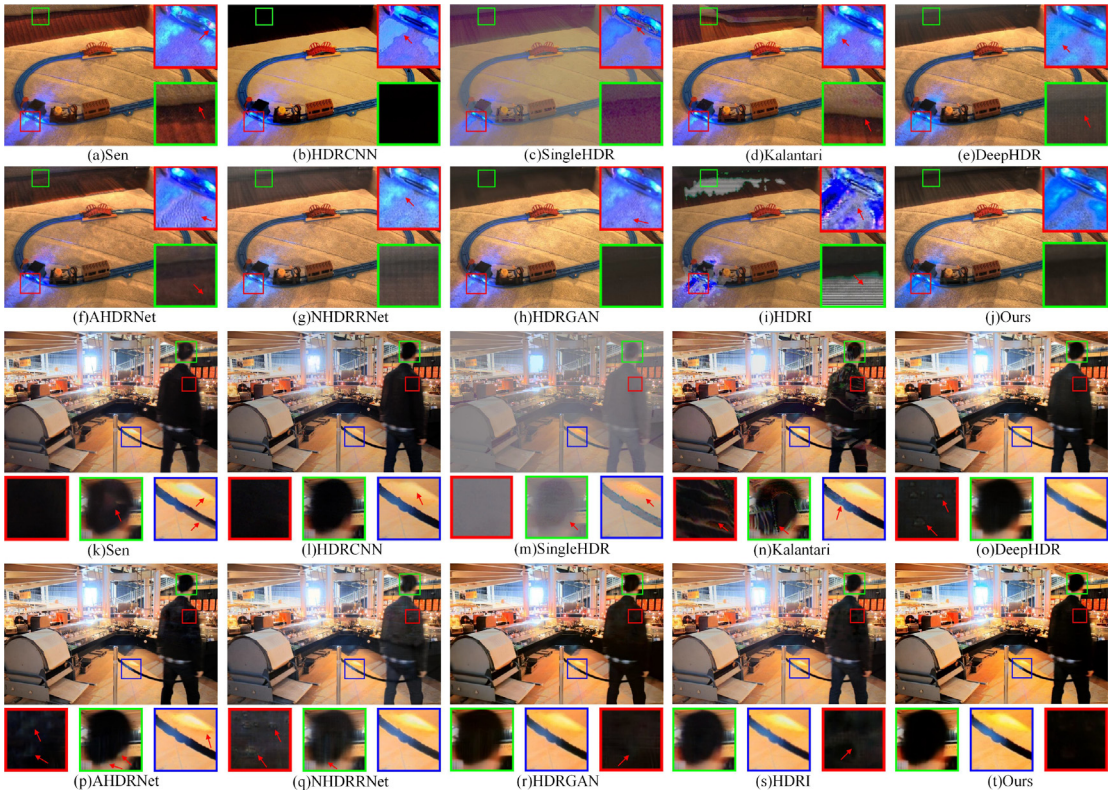

Fig. 9. Comparison of the reconstruction results obtained by different methods on the two sets of LDR images from Tursun's dataset.

Table 1
Performance comparison of the proposed method and state-of-the-art methods on Kalantari's testing dataset.

Methods	PSNR-μ	SSIM-μ	PSNR-L	SSIM-L	HDR-VDP-2
Sen [10]	41.6114	0.9831	40.9453	0.9805	60.4599
HDRCNN [1]	13.8351	0.7800	13.9231	0.4682	51.0721
SingleHDR [6]	12.2975	0.8491	9.6874	0.3358	53.2721
Kalantari [11]	42.7423	0.9877	40.7217	0.9824	63.0420
DeepHDR [12]	41.6377	0.9869	40.8801	0.9857	64.9001
AHDRNet [13]	43.6172	0.9900	41.0390	0.9702	63.8429
NHDRRNet [14]	42.4143	0.9887	37.4557	0.9838	61.2107
HDRGAN [15]	43.8435	0.9907	41.4845	0.9871	64.4497
HDRI [16]	43.1807	0.9892	**41.6862**	0.9864	64.4560
Proposed	**43.9609**	**0.9907**	41.5123	**0.9874**	**65.2112**

color distortion and discrete details in the experimental results. As shown in the enlarged areas marked in Fig. 9(j) and (t), the proposed method truly reconstructs scene structure, clearly recovers image details, and effectively avoid ghosting artifacts.

Quantitative evaluation: Since the samples in Kalantari's testing dataset have the corresponding ground truth, five evaluation metrics PSNR-L, SSIM-L, PSNR-μ, SSIM-μ, and HDR-VDP-2 are used to evaluate the quality of the results obtained by different methods. Table 1 shows the average values of all evaluation results obtained by each method. The proposed method achieves the best performance on PSNR-μ and SSIM-μ, because it not only effectively suppresses ghosting artifacts but also effectively recovers the missing edge details and avoids introducing noise. For HDR-VDP-2, the proposed method also has obvious advantage. The reconstruction results of the proposed method have better visibility and quality. Fig. 10(a) and (b) show the average UDQM and BTMQI values of the reconstruction results obtained by different methods on Tursun's dataset, respectively. Sen's dataset lacks the real exposure time of LDR images, so UDQM cannot be calculated. Only the average BTMQI values of the reconstructed images obtained by different methods on Sen's dataset is shown in Fig. 10(c). According to the results shown in Fig. 10, compared with other methods, the proposed method achieves better performance.

4.5. Ablation study

The proposed method consists of three core modules, CMAN, DMITN, and SEN. In ablation study, the model without these three modules is used as Baseline. After adding channel attention to the Baseline, the model is named as Baseline+CA. After adding CMAN to the Baseline, the model is named as Baseline+CMAN. After adding DMITN to Baseline+CMAN, the model is named as Baseline+CMAN+DMITN. After adding SEN to Baseline+CMAN+DMITN, the model is called Baseline+CMAN+DMITN+SEN. The settings of all experiments in ablation study are the same as those in Section 4.2. Fig. 11 compares the local visual effects of the results obtained by models under different settings when the LDR images shown in Fig. 1 are used to reconstruct the corresponding HDR image.

Effectiveness of CA: As shown in Fig. 11(b), compared with Baseline, Baseline+CA reduces ghosting artifacts to a certain extent. Since CA adjusts the different roles played by different channels of the reference image and other LDR images in HDR image reconstruction, ghosting artifacts are effectively suppressed. The values of objective evaluation indicators shown in Table 2 further verify the effectiveness of CA.

Effectiveness of CMAN: Baseline+CMAN is obtained by adding MA to Baseline+CA. Compared with Fig. 11(b), ghosting artifacts shown in Fig. 11(c) are further weakened. MA can highlight the useful information of the reconstructed image from different levels of the feature maps, so ghosting artifacts are further suppressed. According to objective evaluation metrics shown in Table 2, when CA is replaced by CMAN, the performance of all

Table 2
Quantitative comparison of the performance of different module combinations in the proposed model.

Methods	PSNR-μ	SSIM-μ	PSNR-L	SSIM-L	HDR-VDP-2
Baseline	42.3734	0.9847	40.0215	0.9856	63.4450
Baseline+CA	42.9481	0.9896	40.5967	0.9868	64.3538
Baseline+CMAN	43.3701	0.9903	41.1520	0.9870	64.7209
Baseline+CMAN+DMITN	43.7415	0.9905	41.3672	0.9873	64.8741
Baseline+CMAN+DMITN+SEN	43.9609	0.9907	41.5123	0.9874	65.2112

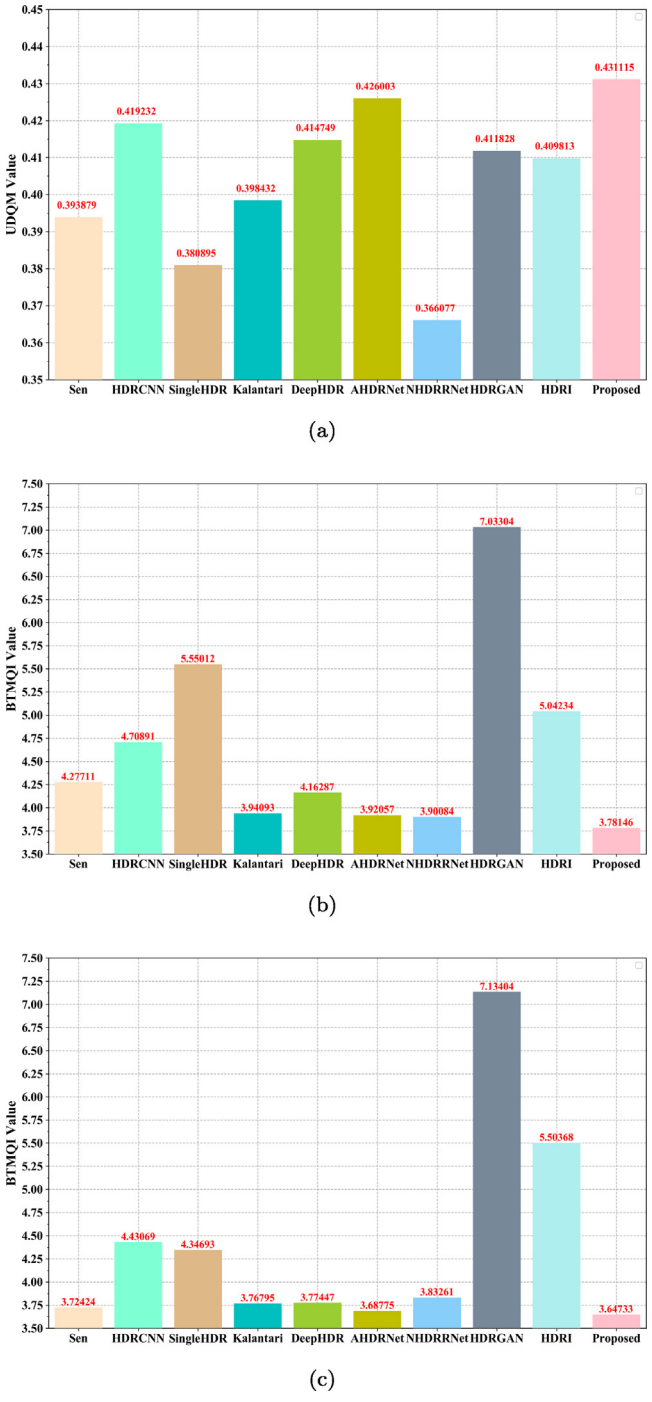

Fig. 10. Comparison of UDQM and BTMQI obtained by different methods on Tursun's and Sen's datasets.

objective evaluation metrics is further improved. Therefore, the effectiveness of CMAN is confirmed.

Table 3
Impact of different values of C' on model performance.

Methods	PSNR-μ	SSIM-μ	PSNR-L	SSIM-L	HDR-VDP-2
C'=0	43.4031	0.9903	41.3565	0.9868	64.5939
C'=4	43.5144	0.9904	41.3206	0.9869	64.7306
C'=6	43.6385	0.9904	41.3953	0.9871	64.9576
C'=8	**43.9609**	**0.9907**	**41.5123**	**0.9874**	**65.2112**
C'=10	43.7171	0.9905	41.4056	0.9873	65.1734
C'=12	43.6826	0.9902	41.2674	0.9868	64.5693

Effectiveness of DMITN: The results obtained by Baseline+CMAN and Baseline+CMAN +DMITN are compared to prove the effectiveness of DMITN. Compared with Baseline+CMAN, ghosting artifacts are effectively suppressed, and local details (e.g. branches) are recovered in the result obtained by Baseline+CMAN+DMITN (as shown in Fig. 11(d)). DMITN can further explore the useful information in the input images to reconstruct HDR image by using different combinations of convolution kernels. According to the values of objective evaluation metrics shown in Table 2, compared with Baseline+CMAN, the overall performance of Baseline+CMAN+DMITN is improved. Therefore, the effectiveness of DMITN is proved.

Effectiveness of SEN: SEN is mainly used to recover the lost details. After SEN is added to Baseline+CMAN+DMITN, Baseline+CMAN+DMITN+SEN improves obviousness and richness of the reconstructed image details. Since SEN can predict the missing details based on the edge structure in the reference image, it not only alleviates the difficulty of recovering lost details in HDR image reconstruction, but also suppresses the negative influence of ghosting artifacts on visual effects. Similarly, the results of objective evaluation metrics shown in Table 2 also prove the effectiveness of SEN.

The impact of the number of heads used in CMAN: The proposed method uses MA to explore the key information of features from different levels. The impact of the number of heads C' on performance is analyzed. The performance of the proposed method is analyzed on Kalantari's dataset when C' is set to 0, 4, 6, 8, 10, and 12. According to Table 3, when the value of C' increases from 4 to 8, the values of most evaluation metrics increase. The values of all evaluation metrics decrease when C' increases from 8 to 12. Therefore, C' is set to 8 in all experiments of this study.

The impact of batch size on model performance: In this section, ablation experiments are performed to select an appropriate batch size. In the experiments, the batch size is set from 1 to 5. The models with different batch size are tested on Kalantari's testing dataset. Fig. 12 show the impact of batch size on the model performance. As shown in Fig. 12, when batch size is 3, the performance of the model is optimal. Therefore, batch size is set to 3.

4.6. Parameter selection and analysis

In Eq. (24), the values of two hyperparameters α and β need to be determined. They are analyzed on Kalantari's testing dataset. On the premise of fixing one parameter, another parameter is adjusted by manual search. α and β are set to 0.01, 0.05, 0.1, 0.5, and 1, respectively. Fig. 13 illustrates the PSNR-L/SSIM-L

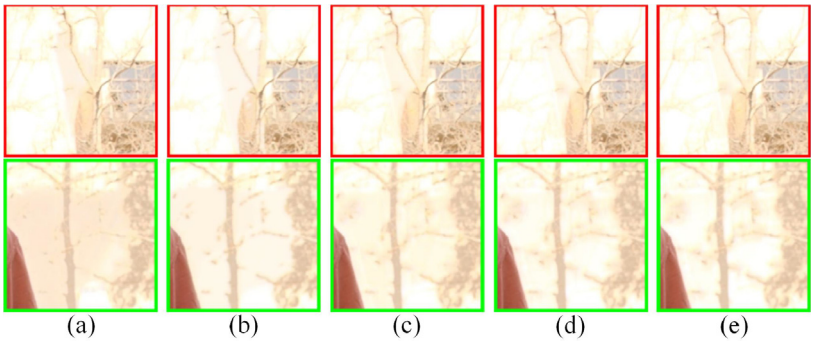

Fig. 11. Influence of each module in the proposed method on the visual performance of the reconstructed HDR images. (a) Partial effect of the result obtained by Baseline, (b) Partial effect of the result obtained by Baseline+CA, (c) Partial effect of the result obtained by Baseline+CMAN, (d) Partial effect of the result obtained by Baseline+CMAN+DMITN, (e) Partial effect of the result obtained by Baseline+CMAN+DMITN+SEN.

Table 4
Performance comparison of the proposed method and state-of-the-art methods on the number of model parameters and the inference time.

Methods	Sen	HDRCNN	SingleHDR	Kalantari	DeepHDR	AHDRNet	NHDRRNet	HDRGAN	HDRI	Proposed
Params(M)	–	29.40	83.30	0.30	20.40	1.24	38.10	2.56	6.70	0.89
Time (s)	61.81	0.44	13.08	29.14	0.28	0.30	0.31	0.69	20.61	0.53

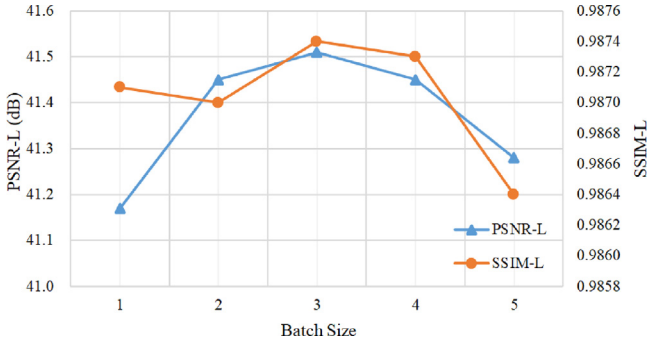

Fig. 12. Impact of batch size on model performance.

curves when each hyperparameter changes. As shown in Fig. 13, when $\alpha = 0.5$ and $\beta = 0.1$, PSNR-L and SSIM-L obtain the optimal values. Therefore, this study sets α and β to 0.5 and 0.1, respectively.

In addition, the influence of the calibration coefficient γ in Eq. (1) on model performance is analyzed. The models with different values of γ are tested on Kalantari's testing dataset. Fig. 14 shows the PSNR-L/SSIM-L curves when γ changes. According to Fig. 14, when γ is 2.2, the model achieves the best performance. Therefore, this study sets $\gamma = 2.2$.

4.7. Time cost and parameter amount

This section compares the time cost and parameter amount of different methods. Table 4 shows the corresponding results. In Table 4, "-" means no correlation, "Time" means the inference time, and "Params" means the model's parameter amount. The inference time is the average time to process images from 15 scenes in Kalantari's testing dataset.

The block match-based Sen's method consumed much more time than the other methods. In neural network-based methods, Kalantari's method took more time due to preprocessing by optical flow. The proposed method took moderate inference time. However, it uses fewer model parameters and has better fusion performance at the same time consumption level. Therefore, such results are acceptable. In general, the proposed method balances the inference time and the number of model parameters well, and achieves better visual performance and quantitative results.

5. Conclusion

This study proposes a structure-embedded ghosting artifact suppression model consisting of three modules, CMAN, DMITN, and SEN, to achieve HDR image reconstruction from multiple LDR images involving the target motion. The proposed model not only solves the information loss caused by the occlusion and motion of targets but also effectively suppresses ghosting artifacts generated in HDR image reconstruction. The designed CMAN can highlight useful information about the reconstructed image, which is conducive to ghosting artifact suppression and recovering lost details in saturated areas. The proposed DMITN ensures the richness of feature information and provides the extracted features with a strong characterization ability, which plays a positive role in ghosting artifact suppression. In addition, because the proposed SEN can predict the edge details of the reconstructed HDR image from the detailed edge information of the reference image, it not only preserves the existing edge details but also achieves the recovery of missing detailed information. The proposed method is superior to state-of-the-art methods on three public HDR datasets.

CRediT authorship contribution statement

Lingfeng Tang: Methodology, Investigation, Software, Data curation, Writing – original draft. **Huan Huang:** Methodology, Writing – original draft. **Yafei Zhang:** Conceptualization, Supervision, Writing – reviewing and editing, Funding acquisition. **Guanqiu Qi:** Writing – reviewing and editing, Validation, Formal analysis. **Zhengtao Yu:** Writing – review and editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Public datasets are used.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 62161015).

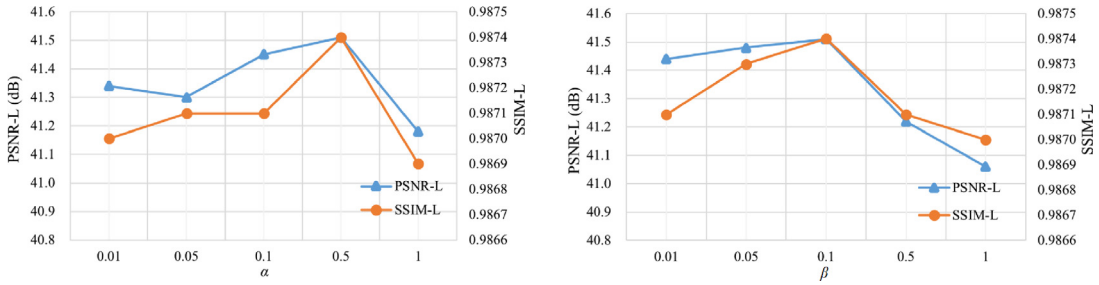

Fig. 13. Hyperparameter analysis on Kalantari's testing dataset. PSNR-L and SSIM-L curves with α and β.

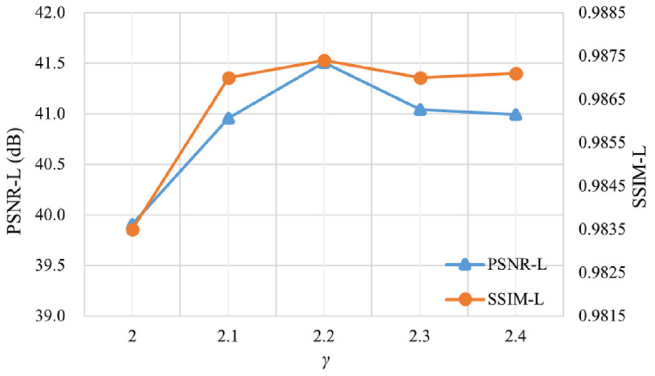

Fig. 14. Impact of the parameter γ on model performance.

References

[1] G. Eilertsen, J. Kronander, G. Denes, R.K. Mantiuk, J. Unger, HDR image reconstruction from a single exposure using deep CNNs, ACM Trans. Graph. 36 (6) (2017) 1–15.

[2] S. Lee, S.Y. Jo, G.H. An, S.-J. Kang, Learning to generate multi-exposure stacks with cycle consistency for high dynamic range imaging, IEEE Trans. Multimed. 23 (2020) 2561–2574.

[3] X. Chen, Y. Liu, Z. Zhang, Y. Qiao, C. Dong, Hdrunet: Single image HDR reconstruction with denoising and dequantization, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2021, pp. 354–363.

[4] K. Fotiadou, G. Tsagkatakis, P. Tsakalides, Snapshot high dynamic range imaging via sparse representations and feature learning, IEEE Trans. Multimed. 22 (3) (2020) 688–703.

[5] G. Chen, L. Zhang, M. Sun, Y. Gao, P.N. Michelini, Y. Wu, Single-image hdr reconstruction with task-specific network based on channel adaptive RDN, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2021, pp. 398–403.

[6] Y.-L. Liu, W.-S. Lai, Y.-S. Chen, Y.-L. Kao, M.-H. Yang, Y.-Y. Chuang, J.-B. Huang, Single-image HDR reconstruction by learning to reverse the camera pipeline, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 1651–1660.

[7] W. Hu, M. Seifi, E. Reinhard, Over-and under-exposure reconstruction of a single plenoptic capture, ACM Trans. Multimed. Comput. Commun. Appl. 14 (2) (2018) 1–21.

[8] F. Kou, Z. Wei, W. Chen, X. Wu, C. Wen, Z. Li, Intelligent detail enhancement for exposure fusion, IEEE Trans. Multimed. 20 (2) (2018) 484–495.

[9] J.-L. Yin, B.-H. Chen, Y.-T. Pengg, Two exposure fusion using prior-aware generative adversarial network, IEEE Trans. Multimed. 24 (2022) 2841–2851.

[10] P. Sen, N.K. Kalantari, M. Yaesoubi, S. Darabi, D.B. Goldman, E. Shechtman, Robust patch-based HDR reconstruction of dynamic scenes, ACM Trans. Graph. 31 (6) (2012) 1–11.

[11] N.K. Kalantari, R. Ramamoorthi, Deep high dynamic range imaging of dynamic scenes, ACM Trans. Graph. 36 (4) (2017) 1–12.

[12] S. Wu, J. Xu, Y.-W. Tai, C.-K. Tang, Deep high dynamic range imaging with large foreground motions, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 120–135.

[13] Q. Yan, D. Gong, Q. Shi, A.v.d. Hengel, C. Shen, I. Reid, Y. Zhang, Attention-guided network for ghost-free high dynamic range imaging, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 1751–1760.

[14] Q. Yan, L. Zhang, Y. Liu, Y. Zhu, J. Sun, Q. Shi, Y. Zhang, Deep HDR imaging via a non-local network, IEEE Trans. Image Process. 29 (2020) 4308–4322.

[15] Y. Niu, J. Wu, W. Liu, W. Guo, R.W. Lau, HDR-gan: HDR image reconstruction from multi-exposed LDR images with large motions, IEEE Trans. Image Process. 30 (2021) 3885–3896.

[16] H. Chung, N.I. Cho, High dynamic range imaging of dynamic scenes with saturation compensation but without explicit motion compensation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2951–2961.

[17] H. Zimmer, A. Bruhn, J. Weickert, Freehand HDR imaging of moving scenes with simultaneous resolution enhancement, Comput. Graph. Forum 30 (2) (2011) 405–414.

[18] Y.S. Heo, K.M. Lee, S.U. Lee, Y. Moon, J. Cha, Ghost-free high dynamic range imaging, in: Asian Conference on Computer Vision, ACCV, 2010, pp. 486–500.

[19] T. Jinno, M. Okuda, Motion blur free HDR image acquisition using multiple exposures, in: 15th IEEE International Conference on Image Processing, ICIP, 2008, pp. 1304–1307.

[20] J. Zheng, Z. Li, Z. Zhu, S. Wu, S. Rahardja, Hybrid patching for a sequence of differently exposed images with moving objects, IEEE Trans. Image Process. 22 (12) (2013) 5190–5201.

[21] H. Li, Y. Cen, Y. Liu, X. Chen, Z. Yu, Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion, IEEE Trans. Image Process. 30 (2021) 4070–4083.

[22] Y. Liu, L. Wang, J. Cheng, C. Li, X. Chen, Multi-focus image fusion: A survey of the state of the art, Inf. Fusion 64 (2020) 71–91.

[23] H. Li, J. Gao, Y. Zhang, M. Xie, Z. Yu, Haze transfer and feature aggregation network for real-world single image dehazing, Knowl.-Based Syst. 251 (2022) 109309.

[24] H. Li, N. Dong, Z. Yu, D. Tao, G. Qi, Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification, IEEE Trans. Circuits Syst. Video Technol. 32 (5) (2021) 2814–2830.

[25] S. Li, F. Li, K. Wang, G. Qi, H. Li, Mutual prediction learning and mixed viewpoints for unsupervised-domain adaptation person re-identification on blockchain, Simul. Model. Pract. Theory 119 (2022) 102568.

[26] H. Li, Y. Chen, D. Tao, Z. Yu, G. Qi, Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification, IEEE Trans. Inf. Forensics Secur. 16 (2020) 1480–1494.

[27] H. Li, K. Xu, J. Li, Z. Yu, Dual-stream reciprocal disentanglement learning for domain adaptation person re-identification, Knowl.-Based Syst. 251 (2022) 109315.

[28] J. Johnston, Y. Li, M. Lops, X. Wang, ADMM-net for communication interference removal in stepped-frequency radar, IEEE Trans. Signal Process. 69 (2021) 2818–2832.

[29] Z. Chen, J. Tang, X.Y. Zhang, Q. Wu, D.K.C.S. Yuxin Wang, S. Jin, K.-K. Wong, Offset learning based channel estimation for intelligent reflecting surface-assisted indoor communication, IEEE J. Sel. Top. Sign. Proces. 16 (1) (2022) 41–55.

[30] Q. Yan, D. Gong, P. Zhang, Q. Shi, J. Sun, I. Reid, Y. Zhang, Multi-scale dense networks for deep high dynamic range imaging, in: 2019 IEEE Winter Conference on Applications of Computer Vision, WACV, 2019, pp. 41–50.

[31] Q. Yan, B. Wang, L. Zhang, J. Zhang, Z. You, Q. Shi, Y. Zhang, Towards accurate HDR imaging with learning generator constraints, Neurocomputing 428 (2021) 79–91.

[32] G. Ward, Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures, J. Graph. Tools 8 (2) (2003) 17–30.

[33] A. Tomaszewska, R. Mantiuk, Image registration for multi-exposure high dynamic range image acquisition, in: The 15th International Conference in Central Europe on Computer Graphics, WSCG, 2007, pp. 49–56.

[34] J. Hu, O. Gallo, K. Pulli, X. Sun, HDR deghosting: How to deal with saturation? in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1163–1170.

[35] R. Feng, H. Shen, J. Bai, X. Li, Advances and Opportunities in Remote Sensing Image Geometric Registration: A systematic review of state-of-the-art approaches and future research directions, IEEE Geosci. Remote Sens. Mag. 9 (4) (2021) 120–142.
[36] J. Xiong, Y. Luo, G. Tang, An improved optical flow method for image registration with large-scale movements, Acta Automat. Sinica 34 (7) (2008) 760–764.
[37] R. Feng, X. Li, H. Shen, Mountainous remote sensing images registration based on improved optical flow estimation, in: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. IV-2/W5, 2019, pp. 479–484.
[38] M. Gong, S. Zhao, L. Jiao, D. Tian, S. Wang, A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information, IEEE Trans. Geosci. Remote Sens. 52 (7) (2014) 4328–4338.
[39] H. Goncalves, L. Corte-Real, J.A. Goncalves, Automatic image registration through image segmentation and SIFT, IEEE Trans. Geosci. Remote Sens. 49 (7) (2011) 2589–2600.
[40] R. Feng, Q. Du, X. Li, H. Shen, Robust registration for remote sensing images by combining and localizing feature- and area-based methods, ISPRS J. Photogramm. Remote Sens. 151 (2019) 15–26.
[41] O. Gallo, N. Gelfandz, W.-C. Chen, M. Tico, K. Pulli, Artifact-free high dynamic range imaging, in: IEEE International Conference on Computational Photography, ICCP, 2009, pp. 1–7.
[42] S. Raman, S. Chaudhuri, Reconstruction of high contrast images for dynamic scenes, Vis. Comput. 27 (12) (2011) 1099–1114.
[43] K.R. Prabhakar, S. Agrawal, D.K. Singh, B. Ashwath, R.V. Babu, Towards practical and efficient high-resolution HDR deghosting with CNN, in: Proceedings of the European Conference on Computer Vision, ECCV, 2020, pp. 497–513.
[44] E. Reinhard, G. Ward, S. Pattanaik, High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting, Morgan Kaufmann Publishers Inc., 2005.
[45] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[46] O.T. Tursun, A.O. Akyüz, A. Erdem, E. Erdem, An objective deghosting quality metric for HDR images, Comput. Graph. Forum 35 (2) (2016) 139–152.
[47] D.P. Kingma, J.L. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, ICLR, 2015, pp. 1–15.
[48] R. Mantiuk, K.J. Kim, A.G. Rempel, W. Heidrich, HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions, ACM Trans. Graph. 30 (4) (2011) 1–14.
[49] K. Gu, S. Wang, G. Zhai, S. Ma, X. Yang, W. Lin, W. Zhang, W. Gao, Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure, IEEE Trans. Multimed. 18 (3) (2016) 432–443.