Fast and flexible stack-based inverse tone mapping

Zhang, Ning; Ye, Yuyao; Zhao, Yang; Li, Xufeng; Wang, Ronggang

Published in:
CAAI Transactions on Intelligence Technology

Published: 01/12/2023

Document Version:
Final Published version, also known as Publisher’s PDF, Publisher’s Final version or Version of Record

License:
CC BY-NC-ND

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1049/cit2.12188

Publication details:

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher’s definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher’s copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
ORIGİNAL RESEARCH

Fast and flexible stack-based inverse tone mapping

Ning Zhang¹ | Yuyao Ye¹ | Yang Zhao² | Xufeng Li³ | Ronggang Wang¹

¹School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen, China
²School of Computer and Information, Hefei University of Technology, Hefei, China
³Department of Computer Science (CS), College of Engineering (EG), City University of Hong Kong, Hong Kong, China

Correspondence
Ronggang Wang, School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Nanshan, China.
Email: rgwang@pkusz.edu.cn

Funding information
Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents, Grant/Award Number: RCJC2020714114435057; National Natural Science Foundation of China, Grant/Award Numbers: 62072013, U21B2012; Shenzhen Science and Technology Program-Shenzhen Hong Kong joint funding project, Grant/Award Number: SGDX202112314490001; Outstanding Talents Training Fund in Shenzhen

Abstract
Inverse tone mapping technique is widely used to restore the lost textures from a single low dynamic range image. Recently, many stack-based deep inverse tone mapping networks have achieved impressive results by estimating a set of multi-exposure images from a single low dynamic range input. However, there are still some limitations. On the one hand, these methods usually set a fixed length for the estimated multi-exposure stack, which may introduce computational redundancy or cause inaccurate results. On the other hand, they neglect that the difficulties of estimating each exposure value are different and use the identical model to increase or decrease exposure value. To solve these problems, the authors design an exposure decision network to adaptively determine the number of times the exposure of low dynamic range input should be increased or decreased. Meanwhile, the authors decouple the increasing/decreasing process into two sub-modules, exposure adjustment and optional detail recovery, based on the characteristics of different variations of exposure values. With these improvements, this method can fast and flexibly estimate the multi-exposure stack from a single low dynamic range image. Experiments on several datasets demonstrate the advantages of the proposed method compared to state-of-the-art inverse tone mapping methods.

KEYWORDS
2-D, image enhancement, image processing

1 | INTRODUCTION

Compared with traditional low dynamic range (LDR) images, high dynamic range (HDR) images can provide a more realistic visual experience and richer details [1]. Currently, the HDR content is usually shot directly by professional devices or generated by multi-exposure fusion [2, 3]. However, there exist a large amount of single LDR image and video resources. Therefore, how to recover the missing details from the single LDR image has attracted wide attention.

Converting a single LDR image to an HDR radiance is also called inverse tone mapping (ITM). Banterle et al. [4] extended the range in the high luminance areas using an inverse Photographic Tone Reproduction operator. Rempel et al. [5] proposed a video ITM method that can deal with video streams in real-time. Akyüz et al. [6] proposed that simply boosting the range of an LDR image linearly is good enough. However, these traditional methods cannot recover realistic textures in the over/under-exposed regions. Recently, researchers have shown an increased interest in deep-learning-based ITM methods, which can be roughly divided into two categories, that is, direct mapping and stack-based methods. The intuitive idea is directly learning an end-to-end model to recover the 32-bits HDR image from the single 8-bits LDR input [7, 8]. However, on the one hand, the change of HDR pixels is more complicated than that of LDR images. On the other hand, the dynamic ranges of HDR images tend to vary greatly from scene to scene, which increases the difficulty of training networks for such an ill-posed problem. The other category is to first estimate the multi-exposure stack (MES) by increasing/decreasing the exposure value (EV) of the input single LDR image and then synthesise them into the HDR radiance [9–11] by multi-exposure fusion technology [12–15]. Consequently, these stack-based methods can effectively...
reduce the learning difficulty because they only need to learn the relative change of pixels in the LDR domain instead of crossing the LDR and HDR domains [10]. However, the stack-based approaches tend to be more time-consuming compared to the direct mapping approaches, which affects their performance in practical applications.

We analyse the limitations of the previous stack-based methods and propose a fast and flexible deep stack-based ITM method. The previous stack-based methods suffer from the fixed length of the MES, which may involve unnecessary computational cost if a smaller length can already restore the dynamic range of the LDR input or cause inadequate dynamic ranges if the fixed length is not enough to recover the lost details. To overcome this limitation, we propose an exposure decision model (EDM) to determine in advance how to estimate the MES. For the input LDR image, the EDM can decide the number of times that the input LDR image should be increased or decreased. In this way, the MES can be generated more appropriately with less redundancy. In addition, different from previous methods that treat the changes between each EV equally, we decouple increasing/decreasing the EV of the single LDR image to the exposure adjustment and optional detail restoration. The decoupling is based on that the difficulties of increasing or decreasing EVs are very different, which depend on whether texture information is available. Specifically, we firstly design a lightweight exposure adjustment model (EAM) which can adjust the luminance, colour, and contrast of the image based on the available information. Then, for the challenging task of recovering the lost details in the over/under-exposed regions, we utilise a complicated but optional detail restoration model (DRM) to regain these textures separately. Figure 1 demonstrates that our method can recover lost details of the LDR images and generate pleasing results.

This paper has the following contributions:

1. The previous stack-based methods set a fixed length for MES, which may lead to inaccurate results and introduce computational redundancy. An EDM is proposed to overcome this limitation, which can flexibly estimate the times of increasing and decreasing according to the specific lighting conditions of the input LDR image, thus reducing computational cost and improving the performance.

2. We divide the generation of the MES into exposure adjustment and optional detail restoration based on the different characteristics of changes between each EV, which further reduces the redundant parameters and computations significantly.

3. By decoupling the complicated EV changing process, each region of the input LDR image can be processed more accurately. The proposed EDM can adjust the entire luminance while avoiding artefacts and the DRM can focus on recovering the lost textures to generate a more realistic result.

2 | RELATED WORK

2.1 | Deep-learning-based inverse tone mapping

Eilertsen et al. [7] focussed on the over-exposed regions and for the other under-exposed regions they used a fixed inverse camera response function (CRF) to roughly process. Santo et al. [16] took the same strategy as Ref. [7] but they used the attention masks to help the model focus on the available features to restore the lost details in the over-exposed regions. Marnerides et al. [8] designed a three-branches model to expand the dynamic range of the single LDR image and avoided artefacts caused by the U-Net [17]. Liu et al. [18] directly reconstructed HDR images by modelling the inverse process of the image formation pipeline, which consisted of dequantisation that converted the 8-bits LDR image to 16-bits, linearisation that transferred the non-linear image into linear radiance, and hallucination that recovered the clipped information. Zheng et al. [19] proposed an ultra-high-definition HDR reconstruction method via a collaborative learning manner that learnt the content and colour details in the dual-path network. Marnerides et al. [20] presented a generative adversarial network-based method that hallucinated missing

**FIGURE 1** The reconstructed HDR images by the proposed method. The lost details can be recovered by the proposed exposure adaptive stack-based ITM method. All the reconstructed HDR images have been tone-mapped by [45] for display.
information from badly exposed areas in LDR images. Chen et al. [21] used a spatially dynamic encoder–decoder network to learn an end-to-end mapping for single image HDR reconstruction with denoising and dequantisation.

Different from the direct mapping, Endo et al. [9] firstly proposed a stack-based method by generating LDR images with different EVs and synthesising them into the HDR images. Specifically, they designed a 2D-encoder with 3D-decoder architecture to estimate a set of exposure increased or decreased images at once, which greatly increases the computational cost due to 3D convolutions. Lee et al. [10] devised a chain structure to generate the increased/decreased images in sequence. Both of them can only estimate MES with fixed length. Once you need to further increase or decrease the EV of the LDR image, a new model must be retrained. Lee et al. [11] incorporated adversarial loss during training the network, which made the generated MES closer to the real one perceptually. Meanwhile, they overcame the above problem by training a single increasing/decreasing model and reusing it to change the EV. Kim et al. [22] proposed a novel framework with a fully differentiable HDR imaging process by introducing the differentiable HDR synthesis layer. Compared to the previous stack-based ITM methods, the proposed method can generate the multi exposure stack with flexible exposure setting and achieve more accurate results.

### 2.2 Deep-learning-based over/under-exposed correction

Increasing EV can be regarded as an under-exposed correction process, which has been studied for a long time. Lore et al. [23] proposed a stacked auto-encoder to learn the natural low-light image enhancement. Wei et al. [24] designed an end-to-end deep network to enhance the low-light images based on the Retinex theory. Li et al. [25] proposed the robust Retinex model, which additionally considered a noise map compared with the conventional Retinex model and thus improved the performance of enhancing low-light images accompanied by noises. Jiang et al. [26] proposed an effective unsupervised generative adversarial network for low-light enhancement, which can be trained without low/normal-light image pairs. Guo et al. [27] proposed a zero-shot learning method to enhance the low light images by estimating curve parameters. The estimation network was lightweight but efficient due to the intuitive and simple non-linear curve mapping. Similarly, decreasing EV also involves the over-exposed correction. Zhang et al. [28] proposed a dual illumination estimation to simultaneously process under-exposure and over-exposure images. Yang et al. [29] corrected the exposure by transferring the image into the HDR domain to restore the information and converting it back. Cao et al. [30] recovered the lost detail in the over-exposed region by disentangling the exposure and scene information. Compared to the deep-learning-based over/under-exposed correction methods, the proposed method aims to recover both the lost details in the over and under-exposed regions and divide the generation of the result into exposure adjustment and optional detail restoration based on the different characteristics.

### 2.3 Exposure bracketing selection

The exposure bracketing selection aims to select the proper sets of exposures to represent the entire information of the scenes. Beek et al. [31] proposed irradiance distribution adaptive bracketing selection algorithm by formulating the selection problem as a polynomially solvable problem. Wang et al. [32] designed an agent and used reinforcement learning (RL) to solve the exposure bracketing selection problem. Specifically, they calculated the peak signal to noise ratio (PSNR) between the possible exposure sets and the ground truth stack, then the PSNR was used as the reward to train the agent. However, on the one hand, it is unnecessary to use RL to do this task because there is no interaction between the agent and the environment. On the other hand, the PSNR metric cannot evaluate the quality of HDR images precisely. Meanwhile, compared to the previous exposure bracketing selection methods, the proposed EDM predict the relationship between each exposure combination to achieve more precise results.

### 3 METHODOLOGY

At first, the limitations of the previous stack-based ITM algorithms are analysed. Then, the proposed method is introduced to solve these problems.

#### 3.1 Problems of stack-based deep ITM methods

The purpose of the stack-based ITM method is to restore the realistic textures in different dynamic ranges by estimating images with different EVs, which can be formulated as:

$$ O_1, O_2, ..., O_n = f(I), $$

where $I$ is the input single LDR image, $O_{t\in[1,n]}$, denotes the predicted $n$ LDR images with different EVs, and $f(C)$ is the stack-based ITM algorithm. Figure 2 shows an example of MES captured by changing the shutter time of the LDR camera. The stack-based ITM methods aim to estimate the rest of the images in the MES from the input LDR image, that is, Figure 2d. For the input LDR image, previous stack-based ITM methods [9–11] usually generate the MES with a fixed length of $2T + 1$ (e.g. when $T = 3$ there are three exposure-increased images and three exposure-decreased images). However, a fixed number $T$ is not always the reasonable and optimal choice. For example, we only need to decrease the exposure if the input image has already covered the most dynamic range of the original scene and only has over-exposed regions. In these cases, a large $T$ involves many unnecessary computations and may introduce artefacts caused by the
algorithms. Conversely, if the original scene has an extremely high dynamic range, a small \( T \) is not sufficient to recover the desired result. Figure 3 shows an example that contains extremely dark regions (blue rectangle) and slightly over-exposed areas (red rectangles). The previous stack-based ITM method [10] still generates three exposure-decreased and three exposure-increased images for the input LDR image. However, the \(-1\) EV image has already restored the textures in the over-exposed regions and the \(+3\) EV image is still not enough to recover the details in extreme dark regions. On the contrary, the proposed method can estimate the MES flexibly, which only decreases the exposure of the input LDR one time and increases the exposure four times to restore sufficient information.

In addition, there are over/under-exposed regions in Figure 2d, which are marked by yellow and red rectangles separately. Through experiments, we find that a simple model is enough to adjust the EV of the well-exposed regions, that is, the rest regions of Figure 2d. However, it is more difficult when it comes to the under/over-exposed regions. On the one hand, the textures and details are almost lost, for example, some pixels in these regions are equal to 0 or 255, which means that there is little available information. On the other hand, the noises in the dark regions of the input LDR image are also amplified during increasing the EV, which further increases the difficulty. However, most previous methods [9–11] neglect this phenomenon and use a complex model to handle them equally, which causes unnecessary computational cost.

Based on these analyses, we devise an EDM to adaptively select the length of the MES. Meanwhile, we propose a lightweight but efficient EAM and a specific DRM to tackle different EV changes separately. The details of each model are introduced in the following and the structures of the proposed models are shown in Figure 4.

### 3.2 Exposure decision model

The EDM aims to decide the times to increase and decrease the EV of the input LDR image. At first, we should generate the label that indicates the best exposure combination. Specifically, we generate 10 EV-increased images and 10 EV-decreased images (which can cover all of the dynamic range of the images) and form 190 different MES combinations. Then, we synthesise the predicted HDR images by the LDR images in each combination and calculate the HDR-VDP-2.2 scores. We find that most of the combinations with the highest scores (98%) located in the range of \(-5\) to 5 EV. Furthermore, the visual quality of these predicted HDR images is also comparable to the ground-truth images. Therefore, we conclude that \( T = 5 \) can handle the great majority of scenes and recover sufficient details. Meanwhile, Barakat et al. [33] has shown that three images can capture the luminance of a scene adequately in most cases, and the input image should be chosen to keep the realistic information. Therefore, we decide to choose two images from the estimated images to form the final MES. Consequently, the total number of possible exposure combinations are \( C_{10}^2 = 45 \). For every single LDR input in the training dataset, we first generate 5 EV-increased images and 5 EV-decreased images and form 45 different MES combinations. Then we synthesise the predicted HDR images by the LDR images in each combination and then calculate the HDR-VDP-2.2 [34] scores \( V, V \in R^{1 \times 45} \) for all predicted HDR images with the ground truth HDR images as reference. The intuitive idea is taking the combination with the highest score as the one-hot label to train a classification model. However, it will ignore the relationship between each combination and decrease the performance. Therefore, we normalise the scores \( V \) as the ground truth and let EDM to predict the relative values. Specifically, we resize each LDR input \( I \) into \( 224 \times 224 \) and send the over/under-exposed masks (binary
masks with the threshold of 0.95/0.05) together to the EDM to predict the relative values $\hat{V}$:

$$\hat{V} = F_{EDM}(\{I, \text{Mask}_{\text{low}}, \text{Mask}_{\text{High}}\}),$$  \hspace{1cm} (2)$$

where $F_{EDM}$ denotes EDM and $\{\cdot\}$ denotes the concatenation operator. The training loss of the EDM is defined as:

$$L_{EDM} = \|\hat{V} - V\|_1.$$  \hspace{1cm} (3)$$

Then, the combination with the highest predict value will be chosen as the final exposure decision. Motivated by Marnerides et al. and Zhang et al. [8, 35], we design the lightweight local-global dense block (LGDB), which consists of a local branch to extract the fine-grained details and a global branch to extract the high-level information (e.g., entire lighting conditions and scene information). The local feature extractor consists of four $3 \times 3$ convolutional (conv) layers with dense connection [36], which aims to preserve the fine-grained details. Note that there is no max-pool operation and thus avoiding the loss of spatial information. In the global feature block, the extracted local features are firstly down-sampled to $32 \times 32$ by an average-pooling operation, and then there are three $3 \times 3$ conv layers with $2 \times 2$ max-pooling and a 4 × 4 conv layer, which aims to extract the global information, for example, entire lighting conditions and scene information. Finally, the local and global features are fused by element-wise addition. There are two linear layers after the LGDB to predict the relative values.

### 3.3 Lightweight exposure adjustment model

The commonly used U-Net [17] structure of previous ITM methods contains too many parameters and thus causes high computational cost when adjusting the exposure of LDR image without over/under-exposed regions. On the contrary, we adopt the proposed LGDB as the lightweight EAM to perform the exposure adjustment:

$$\hat{O}_{EA} = I + F_{EAM}(I),$$  \hspace{1cm} (4)$$

where $I$ and $\hat{O}_{EA}$ denote the input single LDR image and exposure-adjusted output of the network $F_{EAM}$.

The training losses of the EAM contain the pixel-wise loss:

$$L_{\text{pix}} = \|\hat{O}_{EA} - O\|_1,$$  \hspace{1cm} (5)$$

the cosine similarity loss to ensure colour correctness of the RGB vectors of each pixel:

$$L_{\text{cos}} = 1 - \frac{\hat{O}_{EA} \cdot O}{\|\hat{O}_{EA}\|_2 \|O\|_2},$$  \hspace{1cm} (6)$$

and the histogram loss in [22] to ensure that the generated image has a similar global tone with the target image:

$$L_{\text{hist}} = \frac{1}{L} \sum_{l} \|\text{cnt}_l(\hat{O}_{EA}) - \text{cnt}_l(O)\|_1,$$  \hspace{1cm} (7)$$
where \( O \) denotes the ground truth image, \( L \) denotes the intensity levels, and \( cnt_{t} \) indicates the number of pixels, which has a rounded down intensity \( I \) in the input image. The total loss of EAM can be formulated as:

\[
L_{EAM} = L_{\text{pix}} + \lambda_{\text{cos}}L_{\text{cos}} + \lambda_{\text{bias}}L_{\text{bias}},
\]

where \( \lambda_{\text{cos}} \) and \( \lambda_{\text{bias}} \) are set to 5 and 1 separately.

### 3.4 Complicated but optional detail restoration model

Recuperating the lost details in saturated regions is a challenging task, and the EAM module cannot handle this problem due to the lightweight architecture. Consequently, we design a deeper DRM to pointedly solve this problem, which can be formulated as:

\[
\hat{O}_{DR} = \hat{O}_{EA} + F_{\text{DRM}}(\hat{O}_{EA}),
\]

where \( \hat{O}_{DR} \) denotes the detail restored output of the network \( F_{\text{DRM}} \). The DRM is designed based on the U-Net [17] with five-levels and the interpolation and convolution layers are used in the decoder to avoid checkerboard artefacts as in many works [18, 37, 38]. Furthermore, to enhance the ability of the DRM to remove the noises in dark regions, we add random Poisson–Gaussian noise to the input LDR image during training the increasing DRM, which can be approximated by a heteroscedastic Gaussian with a signal-dependent variance:

\[
\sigma^2(I) = I \cdot \sigma^2 + \sigma^2_c.
\]

We then fuse the \( \hat{O}_{DR} \) and \( \hat{O}_{EA} \) with the following equation:

\[
\hat{O} = \hat{O}_{EA} \cdot (1 - \alpha) + \hat{O}_{DR} \cdot \alpha,
\]

where \( \alpha \) is the soft blending mask indicates the over/under-exposed regions of \( I \) and:

\[
\alpha_{\text{decrease}} = \frac{\max(0, \min_{t}(l_{c}) - \tau_{1})}{1 - \tau_{1}},
\]

\[
\alpha_{\text{increase}} = \frac{\min(0, \min_{t}(l_{c}) - \tau_{2})}{-\tau_{2}}.
\]

The training loss of the DRM is defined as:

\[
L_{\text{DRM}} = \lambda_{d} \left\| \hat{O}_{DR} - \hat{O} \right\|^{2} + \lambda_{p} \left\| \phi_{f}(\hat{O}_{DR}) - \phi_{f}(O) \right\|^{2},
\]

where \( \phi_{f} \) denotes the feature map extracted from the \( j \)th max-pooling layer of the VGG-19 [39] pre-trained on ImageNet. We set the \( j \) to 4. The \( \tau_{1}, \tau_{2}, \lambda_{d}, \) and \( \lambda_{p} \) are set to 0.95, 0.05, 40, and 1 separately.

### 3.5 Inference

As shown in Figure 5, for an input LDR image, we firstly resize it to the size of \( 224 \times 224 \) and send it to the EDM to predict the exposure decision \( \langle A, B \rangle \), where \( A, B \in [-5, 0) \cup (0, 5] \). Then, we can obtain the increasing time \( I \) if \( I = \max(A, B) > 0 \), and the decreasing time \( |D| \) if \( D = \min(A, B) < 0 \). During the exposure increasing/decreasing process, the input image is first determined whether there are under/over-saturated (pixels larger than 243 or smaller than 12) regions. If not, the input image is only processed by the EAM, and if yes, the input image will be sent to the DRM to further restore the lost details. Finally, the MES is generated and can be merged into an HDR image byDebevec and Malik [2] or directly merged into the tone mapped image by Kou et al. [3]. We fuse the MES into HDR radiance with the software Photomatix: https://www.hdrsoft.com/.

### 4 EXPERIMENTAL RESULTS

#### 4.1 Implementation details

We collect a multi-exposure dataset, which consists of [40, 41]. There are 776 MES that cover several common HDR scenes such as indoor, skyline, night, and so forth. Because the previous methods only use stacks with fixed length as the dataset, for a fair comparison, we divide this dataset into two

**FIGURE 5** The whole pipeline to process the input LDR image. For example, if \( \langle A, B \rangle \) is \((-3, 1)\), then the LDR image will be increased one time and decreased three times. If \( \langle A, B \rangle \) is \((2, 5)\), then the LDR image will be increased five times and does not need to be decreased.
categories, one contains 429 stacks with a length of 7 and the other has 347 stacks with varying lengths. The stacks with fixed length are then trained into a training set with 222 stacks, a validation set with 31 stacks, and a testing set named HDR-TEST-7 with 62 stacks. The 347 stacks with varying lengths are also used to evaluate the performance and robustness of the proposed method, named HDR-TEST-N. In addition, two publicly available dataset HDR-EYE [42] and HDR-MEF [43] are also used in our evaluation.

During the training process, we first resize the images in the training dataset to 512 × 512 and augment them by randomly cropping to 64 × 64 for EAM and 384 × 384 for DRM. We resize the training images to 224 × 224 during training the EDM without cropping to keep the entire information. The training images are randomly flipped and rotated in all the above training processes. As for the testing datasets, we fuse all the multi-exposure exposure stacks into HDR images by Photomatix, which is a commonly used HDR tool and can generate more accurate fused images than [2]. The resolution of testing images is set to 512 × 512. Our models are implemented using Pytorch 1.2 and deployed on the Tesla V100 GPU. The training batch sizes are 1, 1, and 16 for EAM, DRM, and EDM separately. The training losses are minimised through an Adam optimiser with an initial learning rate of 0.0001.

### 4.2 Comparisons on the predicted HDR images

The proposed method is compared with seven recent state-of-the-art Convolutional Neural Networks-based approaches: Deep reverse tone mapping operator (DrTMO) [9], Deep recursive high dynamic range imaging (HDRI) [11], Deep Single HDRI [18], Deep Chain HDRI [18], Deep Diff HDRI [22], Deep Mask HDRI [16], and Deep HDR-UNet [21]. The commonly used HDR-VDP-2.2 [34] index is adopted to measure the quality of HDR reconstruction. Both the predicted HDR and ground-truth HDR images have been normalised as in [8, 18]. Specifically, the HDR images are normalised in the display-referenced way where the peak luminance is 1000 cd/m², which represents current commercial HDR display technology. The scaling is done to match the 0.1 and 99.9 percentiles of the predictions with the corresponding percentiles of the HDR test images. The input parameters of HDR-VDP-2.2 [34] evaluation is set as follows: a 34-in. display, a viewing distance of 0.5 m and a 1920 × 1080 resolution. For fair comparisons, we fine-tune these models with the same training dataset. Table 1 shows the average HDR-VDP-2.2 scores on the HDR-TEST-7, HDR-TEST-N, HDR-EYE, and HDR-MEF datasets. The proposed method performs favourably against the state-of-the-art methods on all four datasets. Note that compared to the DrTMO [9] with the stack length of 17 and Deep Recursive HDRI [11] with the stack length of 7, the stack length generated by the proposed method is only 3, which can significantly reduce the complexity of the image fusion algorithms.

Figures 6 and 7 show results of these ITM methods. Deep recursive HDRI [11] is easy to blur images. DrTMO [10] causes checkerboard artefacts due to the de-convolution. Deep Single HDRI [18], Deep chain HDRI [10], and Deep Diff HDRI [16] introduce artefacts in the over-exposed regions and fail to recover realistic textures. On the contrary, the proposed method can achieve impressive results both in the over- and under-exposed regions.

### 4.3 Comparisons on the estimated MES

Furthermore, we also compare the quality of the estimated MES with the above stack-based ITM methods and the baseline UNet [17]. Because the multi-exposure images are LDR images, we use the PSNR and Structural Similarity as the evaluation metrics. The quantitative comparison of the estimated MES with previous stack-based methods on the HDR-TEST-7 dataset is shown in Table 2, which also demonstrates the proposed method can estimate more accurate MES than the previous stack-based ITM methods.
4.4 | Comparisons on running time and parameters

We conduct experiments to compare the proposed method with state-of-the-art ITM algorithms on the running time and parameters. Table 3 shows the average running time of each method processing the input LDR images with the size of 512 × 512, 1920 × 1080, and 3840 × 2160 separately and the total parameters of the models. The proposed method can generate the MES faster than previous stack-based ITM methods with few parameters. Note that the running time of stack-based ITM methods only indicates the time of estimating the MES. We set the length of MES to 7 for the U-Net [17], Deep Chain HDRI [10], Deep Recursive HDRI [11], and Deep Diff HDRI [22] as described in their papers. The time of the once increasing/decreasing process of these methods can be calculated by dividing the running time in Table 3 by six. The average processing times of EAM and DRM are 4.5 and 1.3 separately, which on the one hand means that the complicated DRM does not take up too many computations and thus reduces the complexity; on the other hand, further demonstrates the high efficiency due to the EDM, which makes the selection of EVs more flexible. The running time of the once increasing/decreasing process for EAM and DRM can be calculated in the same way.

4.5 | Ablation studies

At first, we conduct experiments to analyse the effect of each proposed module. As shown Figure 8, EAM can adjust the luminance, colour, and contrast of the image based on the available information without introducing artefacts. However, the lost details in the over/under-exposed regions cannot be recovered by the lightweight EAM. With the DRM, the textures in these regions can be restored as close to ground truth as possible. Note that although the proposed DRM is designed based on U-Net [17] without complex improvements, it focuses on recovering lost details in over/under-exposed regions and performs the restoration in LDR domain, so that it can

FIGURE 6 Visual comparison on LDR input image of the testing dataset

FIGURE 7 Visual comparison on LDR input image of the testing dataset
TABLE 2  Quantitative comparison on the estimated MES with previous stack-based methods

<table>
<thead>
<tr>
<th></th>
<th>PSNR+</th>
<th>SSIM+</th>
<th>PSNR−</th>
<th>SSIM−</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net [17]</td>
<td>20.98</td>
<td>0.8437</td>
<td>30.27</td>
<td>0.8508</td>
</tr>
<tr>
<td>Deep chain [10]</td>
<td>27.09</td>
<td>0.8913</td>
<td>30.86</td>
<td>0.8774</td>
</tr>
<tr>
<td>Deep recursive [11]</td>
<td>25.21</td>
<td>0.8455</td>
<td>28.98</td>
<td>0.8151</td>
</tr>
<tr>
<td>Deep diff [22]</td>
<td>26.37</td>
<td>0.8688</td>
<td>31.44</td>
<td>0.8807</td>
</tr>
<tr>
<td>Proposed</td>
<td>29.06</td>
<td>0.9332</td>
<td>32.69</td>
<td>0.9096</td>
</tr>
</tbody>
</table>

Note: Bold values mean best performance.
Abbreviations: MES, multi-exposure stack; PSNR, Peak Signal to Noise Ratio; SSIM, structural similarity.

TABLE 3  Comparisons on the running time (ms) of processing the LDR images with different resolutions and parameters (million) of the models

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct mapping</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDRCNN [7]</td>
<td>47</td>
<td>280</td>
</tr>
<tr>
<td>Deep single HDRI [18]</td>
<td>102</td>
<td>512</td>
</tr>
<tr>
<td>Stack based</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNet [17]</td>
<td>174</td>
<td>1422</td>
</tr>
<tr>
<td>Deep chain HDRI [10]</td>
<td>59</td>
<td>618</td>
</tr>
<tr>
<td>Deep diff HDRI [22]</td>
<td>288</td>
<td>2106</td>
</tr>
<tr>
<td>Proposed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDM</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>EAM</td>
<td>9</td>
<td>54</td>
</tr>
<tr>
<td>DRM</td>
<td>4</td>
<td>33</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>93</td>
</tr>
</tbody>
</table>

Note: Bold values mean best performance.
Abbreviation: HDRI, high dynamic range imaging.

FIGURE 8  Ablation study of the contribution of each module
TABLE 4  Ablation study of the contribution of each module

<table>
<thead>
<tr>
<th>EAM</th>
<th>DRM</th>
<th>EDM</th>
<th>HDR-VDP-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>66.4308</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>66.6636</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>67.4776</td>
</tr>
</tbody>
</table>

Note: Bold values mean best performance.

TABLE 5  Ablation study of the exposure selection methods

<table>
<thead>
<tr>
<th>Classification</th>
<th>RL</th>
<th>EDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDR-VDP-2.2</td>
<td>65.3281</td>
<td>65.9142</td>
</tr>
</tbody>
</table>

Note: Bold values mean best performance.

FIGURE 9  Examples of exposure selections of the proposed exposure decision model

combination (−4,5) for the input image, which means increasing the input image five times and decreasing it four times to recover sufficient details. On the contrary, if the input images do not have over-exposed regions, as shown at the bottom of Figure 9, a fixed-length model may involve unnecessary computational cost while the EDM chooses a combination (1,2) that only needs to increase the exposure two times and thus saves the computations to decrease the exposure.

5  CONCLUSION

We have proposed a novel pipeline to recover the HDR radiance from a single LDR image by estimating the MES fast and adaptively, which avoids unnecessary computational redundancy by analysing the differences of changes between each EV and tackling them separately. An exposure classification model is also presented to adaptively choose appropriate increasing and decreasing times for each input LDR image. The proposed lightweight EAM can change the luminance of the input image while preserving fine-grained details and avoiding artefacts. The DRM can focus on the lost textures in the over/under-exposed regions and recover realistic results. The proposed exposure adaptive ITM method performs favourably against state-of-the-art methods both quantitatively and qualitatively and has a significant advantage in terms of computational cost and running time.

ACKNOWLEDGEMENTS

This work is financially supported by National Natural Science Foundation of China U21B2012 and 62072013, Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents RCJC202007141144435057, Shenzhen Science and Technology Program-Shenzhen Hong Kong joint funding project of SGDX20211123144400001, and Outstanding Talents Training Fund in Shenzhen.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT

The data sets that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Ning Zhang https://orcid.org/0000-0003-3985-4305

REFERENCES


SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.