# Revisiting the Stack-Based Inverse Tone Mapping

Ning Zhang[1,3], Yuyao Ye[1,3], Yang Zhao[2,3], and Ronggang Wang[1,3]
[1]School of Electronics and Computer Engineering, Peking University
[2]School of Computer and Information, Hefei University of Technology [3]Peng Cheng Laboratory
zhangn77@pku.edu.cn yeyuyao@pku.edu.cn yzhao@hfut.edu.cn rgwang@pkusz.edu.cn

## Abstract

*Current stack-based inverse tone mapping (ITM) methods can recover high dynamic range (HDR) radiance by predicting a set of multi-exposure images from a single low dynamic range image. However, there are still some limitations. On the one hand, these methods estimate a fixed number of images (e.g., three exposure-up and three exposure-down), which may introduce unnecessary computational cost or reconstruct incorrect results. On the other hand, they neglect the connections between the up-exposure and down-exposure models and thus fail to fully excavate effective features. In this paper, we revisit the stack-based ITM approaches and propose a novel method to reconstruct HDR radiance from a single image, which only needs to estimate two exposure images. At first, we design the exposure adaptive block that can adaptively adjust the exposure based on the luminance distribution of the input image. Secondly, we devise the cross-model attention block to connect the exposure adjustment models. Thirdly, we propose an end-to-end ITM pipeline by incorporating the multi-exposure fusion model. Furthermore, we propose and open a multi-exposure dataset that indicates the optimal exposure-up/down levels. Experimental results show that the proposed method outperforms some state-of-the-art methods.*

## 1. Introduction

The luminance distribution in nature spans a wide range, from the starlight ($10^{-5}cd/m^2$) to the direct sunlight ($10^{8}cd/m^2$). The low dynamic range (LDR) devices can not cover the full range of luminance of the real scene, and thus fail to reproduce the realistic visual experience. High dynamic range imaging (HDRI) technology [1] [26] can solve this problem, which takes multiple LDR images of the same scene with different shutter time, and then generates the high dynamic range (HDR) image via the multi-exposure fusion (MEF) method [4] [19] [32]. However, HDRI cannot handle the images that have already been captured, such as a large number of LDR images and videos on the Internet. The inverse tone mapping (ITM) technique is thus designed to recover the HDR radiance from a single LDR image, which is an ill-posed problem because the details in the highlights and shadows are almost lost and difficult to be restored. Fortunately, the development of deep learning [6] [7] provides a solution by learning and predicting the distribution of the lost information from the huge amount of training examples.

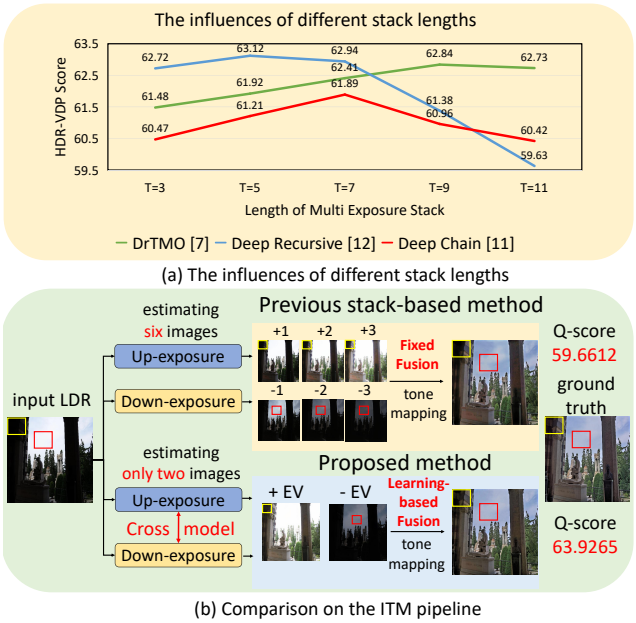


Figure 1. (a) The influences of different stack lengths demonstrate that the stack length is important to the quality of reconstructed HDR image. However, the previous ITM methods [7] [11] [12] [10] simply set a fixed length, which cannot be the optimal choice for every scene. (b) The comparison between the MES predicted by the state-of-the-art stack-based method [10] and the proposed method. Our method only needs to estimate two exposure images to recover more realistic details in highlights and shadows and achieves a higher HDR-VDP-2.2 Q-score. The HDR images are tone mapped by [15] for LDR display.

There are two main deep-learning-based ITM approaches, i.e., direct mapping methods [6] [17] [22] [28] [34] and stack-based methods [7] [10] [11] [12] [13]. The direct mapping methods learn an end-to-end model to recover the HDR radiance from the LDR input straightforwardly. By contrast, the stack-based methods simulate the HDRI technology by increasing and decreasing the exposure value (EV) of the input image to obtain the multi-exposure stack (MES). Compared to the direct mapping methods, the stack-based methods simulate the generation process of HDR images in the real world and perform the learning stage in the same LDR space, which avoids the sophisticated changes between the LDR and HDR domain [10] [11] [13].

The process of the previous stack-based based ITM methods can be roughly summarized in the following three main steps: (1) Training an up-exposure model and a down-exposure model separately and independently. (2) Generating a fixed number of exposure adjusted images (e.g., three up-exposure and three down-exposure images) with the trained models. (3) Merging these images by a classic multi-exposure fusion (MEF) approach [4]. However, each of these three steps has limitations that may cause inaccurate results. Specifically, (1) the process of increasing and decreasing the exposure should not be independent. For instance, when decreasing the exposure, some useful information of the under-exposed regions may become subtle, while the features in the opposite increasing process can compensate for it. (2) The times of exposure adjustment are important to the quality of reconstructed HDR images and a fixed length of MES will cause incomplete information recovery or introduce an unnecessary computational cost, as shown in Fig. 1 (a). (3) The classic MEF approaches cannot be integrated into the entire end-to-end training process. Although Kim et al. [10] proposed a differentiable HDR synthesis layer, it is time-consuming and highly depends on the shutter time and therefore cannot be applied to general scenes.

In this paper, we propose a novel HDR reconstruction method with adaptive exposure adjustment, which provides an effective solution to the existing limitations in the field of stack-based ITM. At first, we design an efficient encoder equipped with the luminance-guided convolution (LGC) and cross-model attention block (CMAB) to extract useful information from local and cross-model features. With the help of CMAB, valid information on the entire up-exposure and down-exposure process can be fully explored to help their reconstruction. Secondly, the proposed up-exposure and down-exposure models can adjust the input LDR image only once to obtain the corresponding optimal exposure adjusted result. In this way, we can avoid the difficulty of determining the length of the MES and get the desired exposure adjustment directly. For this purpose, appropriate ground truth is needed to indicate the optimal exposure level. Therefore, we improve the SICE [2] dataset to form a new MES dataset with optimal exposure labels. On the other hand, since the exposure levels of the labels are different, the models need to be able to generate different results adaptively based on different inputs. Consequently, we devise the exposure adaptive block (EAB) to extract the global information and remap the features of the decoder. The features extracted from EAB are used to normalize the features in the decoder, which results in the image-adaptive capability. Thirdly, we propose a lightweight and fast multi-exposure fusion model (MEFM), which can merge the exposure adjusted results with the input image into the desired HDR image and thus make the whole pipeline end-to-end. Furthermore, we propose progressive reconstruction loss and mask-aware generative adversarial loss to avoid the artifacts in the restored textures of over/under-exposed regions. As Fig. 1 (b) shows, the proposed method only needs to estimate two exposure values to recover the lost information in the shadows and highlights respectively, which is more concise and effective. Experiments show that our ITM algorithm outperforms the state-of-the-art ITM methods in both quantitative and qualitative evaluations.

This paper has the following main contributions:

(1). We propose a novel stack-based ITM framework, which only needs to estimate two exposure images to form the MES. In this way, the lost information can be recovered more efficiently and precisely. Moreover, the exposure adaptive block is designed to adaptively adjust the exposure based on LDR inputs with different luminance distributions.

(2). We connect the up-exposure and down-exposure models with the designed cross-model attention block, which can fully extract the effective features of the image regions with different luminance.

(3). A lightweight and fast multi-exposure fusion network is proposed that can merge the generated results and makes the entire training pipeline end-to-end.

(4). A more concise MES dataset is proposed and opened based on the SICE dataset [2], which contains the optimal exposure-up/down labels to train the adaptive exposure adjustment networks.

## 2. Related Work

**Direct mapping inverse tone mapping.** Eilertsen et al. [6] use the fixed inverse camera response function to linearize the input image and propose an end-to-end network that can predict the lost information in the saturated image areas. Marnerides et al. [22] propose the ExpandNet which uses a multiscale architecture that avoids the use of upsampling layers to improve image quality. Liu et al. [17] incorporate the domain knowledge of the LDR image formation pipeline into their method and learn specialized networks to reverse it. Santo et al. [28] propose attention masks that can

reduce the contribution of useless features in saturated areas. Zheng et al. [34] propose an ultra-high-definition HDR reconstruction method via a collaborative learning manner that learns the content and color details. Chen et al. [3] use a spatially dynamic network to learn an HDR reconstruction with denoising and dequantization.

**Stack-based inverse tone mapping.** Endo et al. [7] propose the first deep stack-based method by estimating LDR images taken with different exposures and reconstructing an HDR image by merging them. Lee et al. [11] propose a stack-based ITM method that produces the MES from the single LDR input using a deep neural network with a chain structure. Lee et al. [12] then reconfigure the deep chain structure by using the generative adversarial network and repeating it recursively to generate the MES. Kim et al. [10] devise the differentiable HDR synthesis layer to replace the conventional fusion method [4] and thus form the end-to-end stack-based ITM network.

**Low-light enhancement and image inpainting.** The increasing exposure of the stack-based ITM methods is similar to the low-light enhancement. Jiang et al. [9] propose the unsupervised learning method and devise the self-regularized attention mechanism. The decreasing exposure aims to recover the lost details of the over-exposed regions, which is similar to the image inpainting that restores the textures in the masked regions. Liu et al. [16] propose the partial convolution which is masked and re-normalized to be conditioned on valid pixels. Yu et al. [33] propose the gated convolution layer which provides a learnable dynamic feature selection mechanism.

# 3. Methodology

## 3.1. Problem formulation

Given a single LDR image $I$, our goal is to predict the exposure-up image $I_{up}$ and the exposure-down image $I_{down}$ with the proposed adaptive up-exposure model (AUEM) and adaptive down-exposure model (ADEM), respectively. Then the predicted images and the input LDR will be merged into the target HDR radiance by the proposed multi-exposure fusion model. The structure of ADEM and AUEM is the same and shown in Fig. 2. We will introduce each component in detail as follows.

## 3.2. Luminance-guided convolution

The aim of the up-exposure or down-exposure model is to adjust the luminance distribution of the input LDR image and recover the lost details in the under-exposed or over-exposed regions. Therefore, the encoder should pay more attention to these regions to extract useful features. We design the luminance-guided convolution (LGC) where the input features are first multiplied by the corresponding luminance-guided map $L$, and then fed to the following convolutional layer, where $L = max(R, G, B)$ in the down-exposure model and $L = 1 - min(R, G, B)$ in the up-exposure model. Note that the "luminance" here is the max/min value of RGB channel instead of the commonly used physical luminance.

## 3.3. Cross-model attention block

As specified in Section 1, the previous stack-based methods neglect the connections between the up-exposure and down-exposure models. To take advantage of useful features in the other model, we design the cross-model attention block (CMAB). Specifically, the input features of the current model $F_{cur}$ and the corresponding features from the same level of the opposite model $F_{ref}$ are fed into two $1 \times 1$ convolutional layers respectively to reduce the channels first.

$$\tilde{F}_q = g(F_q), \tag{1}$$

where $q \in \{cur, ref\}$ and g denotes the $1 \times 1$ convolutional layer. Then $\tilde{F}_{cur}$ and $\tilde{F}_{ref}$ are concatenated into $\tilde{F}_{cat}$ and processed by the another two $1 \times 1$ convolutional layers to calculate the attention scores $M_q$ that indicates which region of the corresponding features is more important. The over/under-exposed regions usually contain very little useful information, e.g., there may be totally saturated pixels for each RGB channel (the pixel values of all three RGB channels are 255). In order to extract long-distance features that may help restore the lost textures, there are four $3 \times 3$ convolutional layers with dilated factors $\{1, 5, 9, 13\}$ respectively to enlarge the receptive fields without introducing extra parameters:

$$F_{d_i} = h_{d_i}(cat(M_{cur} \odot \tilde{F}_{cur}, M_{ref} \odot \tilde{F}_{ref})), \tag{2}$$

where $h_{d_i}$ denotes $3 \times 3$ convolutional layer with dilated factor $d_i$, $cat$ means concatenation along the channel dimension and $\odot$ denotes the element-wise multiplication. Finally, we concatenate the output features $F_{d_i}$ with different dilated factors and perform the channel attention mechanism [31] to exploit the inter-channel relationship of features. Another $1 \times 1$ convolutional layer is added to produce the residual which is added to the input current feature $F_{cur}$:

$$F_{out} = g(CA(cat(F_{d_1}, F_{d_5}, F_{d_9}, F_{d_{13}}))) + F_{cur}, \tag{3}$$

where $CA$ denotes the channel attention [31] and $F_{out}$ means the output features of the proposed CMAB. Fig. 2 (b) illustrates the details of the proposed CMAB.

## 3.4. Exposure adaptive block

The previous stack-based methods only need to learn the relative changes between each EV, which is easier for the model to predict. However, in the proposed pipeline, the luminance and contrast distributions of the input LDR images

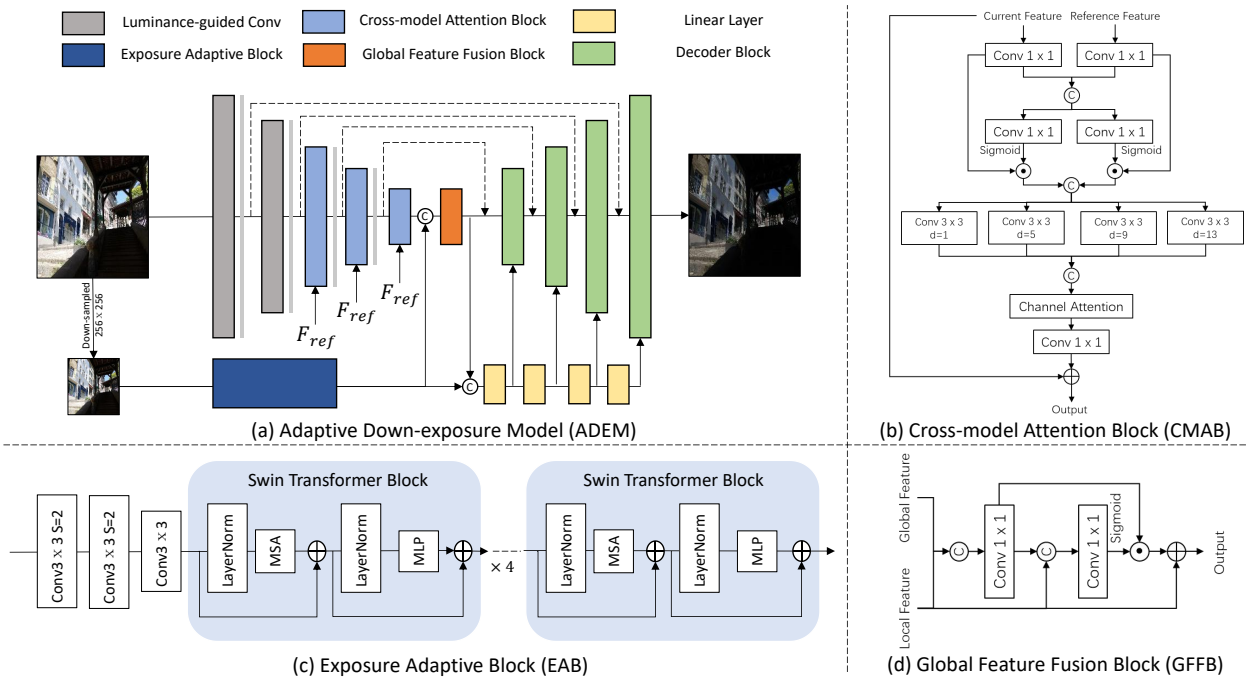


Figure 2. (a) The framework of the proposed adaptive down-exposure model. (b) The details of the proposed cross-model attention block. (c) The details of the proposed exposure adaptive block (the details of the swin transformer block can be found in the supplementary materials). (d) The details of the proposed global feature fusion block.

are different, which leads to a wide variation in the levels of the ground truth exposures. For instance, if the input image contains extremely dark areas, the up-exposure model needs to increase a large EV to recover the textures of the under-exposed regions. On the contrary, if the input image is relatively not so dark, the up-exposure model only needs to increase the EV by a small amount to obtain better details. The commonly used network architecture such as U-Net [27] cannot handle this wide variation, whose weights of convolution kernel are fixed after being trained. In this paper, we design the exposure adaptive block (EAB) to extract the high-level and global information of the input image such as the overall luminance distribution, the scene information, and semantic features. Then we remap the features of the decoder according to the extracted features. Note that the AUEM and ADEM share the same EAB. The input of EAM is the down-sampled LDR image with the resolution of $256 \times 256$. There are three $3 \times 3$ with strides 2, 2, 1 to extract the local features first. Then we incorporate the Swin-Transformer Block [18] [14] into EAB, which has been proven to have impressive performance on modeling the global dependency, to get the global features $F_{glo}$. After that, the $F_{glo}$ is concatenated with the features of the last encoder layer $F_{enc}$ of AUEM or ADEM and processed by the global feature fusion block (GFFB) to get $\tilde{F}_{glo}$, which enriches the global information of the encoded features of AUEM or ADEM. On the other hand, the $F_{glo}$ and $F_{enc}$ are concatenated and processed by the linear layers which map them into the corresponding representations:

$$S_1 = f(cat(AAPool(F_{glo}), AAPool(F_{enc}))), \quad (4)$$

$$S_j = f(S_{j-1}), \quad (5)$$

where f denotes the linear layer, $AAPool$ means an adaptive average pool layer where the resolution of features is pooled to $1 \times 1$, and $S_j$ denotes the output features of the corresponding linear layer. After that we first calculate the mean $\mu_j$ for each individual feature channel of the decoder layers and remap it with the corresponding $S_j$:

$$\tilde{F}_{D_j} = F_{D_j} - \mu_j + S_j, \quad (6)$$

where $F_{D_j}$ is the output feature of the $j-th$ decoder layer and $\tilde{F}_{D_j}$ is the remapped feature. By incorporating the EAB into the AUEM and ADEM, we extract the global features of the input image and use it to control the strength of exposure adjustment to achieve adaptive exposure adjustment. The details of EAB and GFFB are shown in Fig. 2 (c) and (d) separately.

### 3.5. Multi-exposure fusion model

Previous stack-based ITM methods use the classic MEF method [4] to merge the predicted results and the input LDR

into the HDR image. However, it is not differentiable and thus cannnot be incorporated into the end-to-end training process. Although Kim et al. [10] revise it into a differentiable synthesis layer, it is time-consuming due to recovering camera response function by the classic least square method and highly depends on the shutter time. However, the shutter time cannot be accessed in ITM, and an unsuitable shutter time will cause inaccurate results. Meanwhile, although the proposed method can obtain impressive results by estimating only two images that have optimal exposure levels, compared to HDR images reconstructed by the entire MES, using only two exposure-adjusted images may result in that in some scenes with too large dynamic range, there are some areas where the details are not well reconstructed. Therefore, we further design a lightweight but efficient multi-exposure fusion model (MEFM). The ground truth for training MEFM is the HDR images reconstructed with the full MES by the HDR tool Photomatix. In this way, the MEFM can correct the defects caused by estimating only two images and make the whole process end-to-end. The MEFM is based on the ExpandNet [22], which extracts the local, dilated, and global features separately and merges them to generate the result. Instead of predicting the HDR directly, the MEFM estimates three masks $M$ which combine the exposure adjusted images with the input LDR image into $I_{fuse}$:

$$I_{fuse} = M_{up} \odot I_{up} + M_{mid} \odot I + M_{down} \odot I_{down}, \quad (7)$$

and four parameters $a_i, i \in [0, 3]$ which forms a non-linear curve to map the fused LDR into linear HDR radiance $H$:

$$H = a_3 \cdot I_{fuse}^3 + a_2 \cdot I_{fuse}^2 + a_1 \cdot I_{fuse} + a_0. \quad (8)$$

The structure of MEFM is shown in Fig. 3.

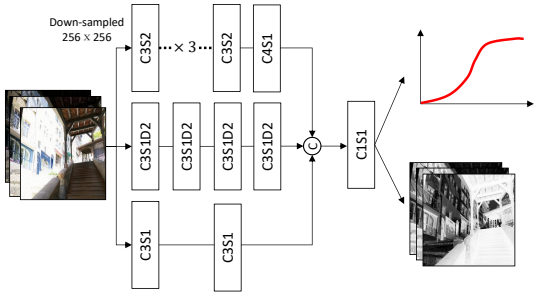


Figure 3. The details of the proposed multi-exposure fusion model. "CxSyDz" denotes convolution layer with kernel size x, stride y and dilation factor z.

### 3.6. Mask-aware discriminator

It is easy to introduce artifacts into the restored textures of over/under-exposed regions. The previous ITM methods [12] [24] introduce GAN to improve results, where they use the discriminator to distinguish the entire real image from the entire fake image. However, the textures in well-exposed areas are easy to adjust and natural enough. Consequently, the discriminator should focus on the under/over-exposed regions to avoid artifacts and achieve more realistic results. Motivated by [29], we design the mask-aware discriminator. The discriminator is based on the 5-levels U-Net [27], which classifies the input images on a global and local per-pixel level. The fake image to be classified is obtained by fusing the under/over-exposed areas of the generated image with the rest of the areas of the ground-truth image through the soft under/over-exposed mask $M_{exp}$. Correspondingly, the pixel-level label of the fake image then becomes $1 - M_{exp}$, which means the mask-aware discriminator needs to judge the under/over-exposed areas as fake and the other areas as true if the input is the mixed fake image. With the mask-aware discriminator, the adversarial loss can effectively make the details of under/over-exposed areas more realistic. The soft under/over-exposed masks are calculated as in [6] [17] and the details of the proposed mask-aware discriminator are shown in Fig. 4.

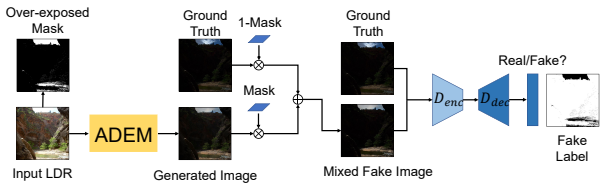


Figure 4. The details of the proposed mask-aware discriminator. The proposed discriminator classifies the input images on per-pixel level, which only focuses on the under/over-exposed regions to achieve more realistic results.

### 3.7. Loss function

The objective loss function contains the progressive reconstruction loss, the mask-aware adversarial loss, and the HDR fusion loss.

**Progressive reconstruction loss.** The reconstruction loss contains the pixel-wise L1 loss between the generated image $I_g$ (i.e., $I_{up}$ and $I_{down}$) and its corresponding ground truth $\hat{I}_g$:

$$\mathcal{L}_{pix}(I_g, \hat{I}_g) = \left| I_g - \hat{I}_g \right|, \quad (9)$$

and the perceptual loss which can force the model to generate images semantically closer to the ground truth:

$$\mathcal{L}_{per}(I_g, \hat{I}_g) = \sum_{l=3}^{5} \left( \left| \phi_l(I_g) - \phi_l(\hat{I}_g) \right| \right) \quad (10)$$

where $\phi_l$ denotes the feature maps extracted by the l-th max-pooling layer of the VGG-16 [30] pre-trained on ImageNet [5]. Directly recovering the textures of over/under-exposed regions is a difficult task. To avoid this problem,

we adopt the progressive generation process and calculate the corresponding reconstruction loss. Specifically, the output features of GFFB $F_{GFFB}$ are mapped into an RGB image $I_{g_1}$ by $1 \times 1$ convolution. After that for each decoder block, there is a $1 \times 1$ convolution which maps the decoded features $\tilde{F}_{D_j}$ into a residual map which is added to the upsampled RGB image predicted by the previous layer:

$$I_{g_j} = g(\tilde{F}_{D_j}) + I_{g_{j-1}} \uparrow . \tag{11}$$

The final progressive reconstruction loss is calculated as:

$$L_{prog}(I_g, \hat{I}_g) = \sum_{j=1}^{5} \mathcal{L}_{pix}(I_{g_j}, \hat{I}_{g_j}) + \lambda \mathcal{L}_{per}(I_{g_j}, \hat{I}_{g_j}). \tag{12}$$

**Adversarial loss.** We adopt the least-square GAN [21] as the adversarial loss. The loss function of the pixel-level decoder of the discriminator $D_d$ are:

$$\mathcal{L}_{D_d} = \frac{1}{2}[(D(\hat{I}_g) - 1)^2] + \frac{1}{2}[(D(z)) - b)^2], \tag{13}$$

where $z = I_g \cdot M_{exp} + \hat{I}_g \cdot (1 - M_{exp})$ is the mixed fake image and $b = 1 - M_{exp}$ is the mixed fake label. Correspondingly, the generator objective becomes:

$$\mathcal{L}_{G_d} = \frac{1}{2}[(D(z) - 1)^2]. \tag{14}$$

**HDR fusion loss.** The classic L1 or L2 is not applicable to the HDR domain, which will make the network focus on high luminance values and underestimate the impact of lower luminance values. Therefore, we calculate the L1 loss $\mathcal{L}_H$ between the tone-mapped images by $\mu$-law as in [17]:

$$\mathcal{L}_H(H, \hat{H}) = \left| \tau(H) - \tau(\hat{H}) \right|, \tag{15}$$

where $\tau(H) = \log(1 + \mu H)/log(1 + \mu)$ and $\hat{H}$ is the HDR image reconstructed with the full MES by the HDR tool Photomatix. Therefore, the total training loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{prog} + \lambda_{gan}(\mathcal{L}_{G_d} + \mathcal{L}_{D_d}) + \lambda_H \mathcal{L}_H, \tag{16}$$

where $\mu$, $\lambda_{per}$, $\lambda_{gan}$, and $\lambda_H$ are set to 5000, 0.5, 0.1, and 1 separately.

# 4. Experimental Results

## 4.1. Implementation details

**Dataset** The SICE-S dataset contains 589 multi-exposure stacks that cover several common HDR scenes such as indoor, skyline, and so on. Because there are some publicly available datasets in SICE-S, e.g., HDR-EYE [23] and HDR-FAIRCHILD [8], we remain them for testing and randomly divide the others into training dataset HDR-TRAIN that contains 360 multi-exposure stacks and testing dataset HDR-TEST that contains 90 multi-exposure stacks. Thus, there are three testing datasets: HDR-TEST, HDR-EYE [23], and HDR-FAIRCHILD [8]. The details of the SICE-S dataset can be found in the supplementary materials.

**Experiment setup** The implementation environment is PyTorch 1.9 version and the Adam optimizer is applied to train the model with the learning rate of 0.0002. We first resize the training pairs in the training dataset to $512 \times 512$ and augment them by randomly cropping to $384 \times 384$. The training images are randomly flipped and rotated. The predicted and ground truth MES in the testing datasets are fused into the HDR images by Photomatix for all of the compared methods with the size of $512 \times 512$.

**Evaluation metrics** The commonly used HDR-VDP-2.2 [20] is adopted to measure the quality of HDR reconstruction. The normalization and parameter settings of evaluation are the same as in [17]. Furthermore, we also evaluate the PSNR and SSIM scores between the tone-mapped LDR images of the ground truth HDR and the predicted HDR by the tone mapping algorithm [15] as in [17].

## 4.2. Comparisons on the predicted HDR images

The proposed method is compared with seven recent state-of-the-art CNN-based approaches: HDRCNN [6], DrTMO [7], Deep Recursive HDRI [12], Deep Single HDRI [17], Deep Synth HDRI [10], Deep Mask [28], and Deep HDRUNet [3]. For fair comparisons, we re-train these models with the same training dataset. For the stack-based methods, the number of predicted images is consistent with the original paper to achieve the best performance. For the direct mapping methods, we use the ground-truth HDR images fused by Photomatrix to train them directly. (Because the previous stack-based ITM methods are not designed for SICE-S, we use the full MES in SICE to train them for better results.)

**Quantitative comparisons.** Table 1 shows the average HDR-VDP-2.2, PSNR, and SSIM scores on the HDR-TEST, HDR-EYE, and HDR-FAIRCHILD datasets. The proposed method performs favorably against the state-of-the-art methods on all three datasets.

**Visual comparisons.** Fig. 5 and Fig. 6 show the results of these ITM methods on two LDR images with significantly different exposures. The HDRCNN [6] and Deep Mask [28] may cause over-bright images because of the use of a fixed camera response function. The DrTMO [7] introduces checkerboard artifacts due to the de-convolution [25]. The Deep Recursive HDRI [12] is easy to blur the images. The Deep Single HDRI [17] and Deep HDRUNet [3] introduce halo artifacts in the over-exposed regions, and the Deep Synth HDRI [10] fails to reconstruct the correct luminance distribution. On the contrary, the proposed method can recover pleasure results both in the under- and over-

Table 1. Quantitative comparison on HDR images with existing methods. Because the training code of Deep Mask [28] is not released, we use the pre-trained model for comparison.

| | HDR-VDP-2.2 | | | TM-PSNR/SSIM | | |
|---|---|---|---|---|---|---|
| | **TEST [2]** | **EYE [23]** | **FAIR [8]** | **TEST [2]** | **EYE [23]** | **FAIR [8]** |
| **HDRCNN [6]** | 63.27 | 55.14 | 58.15 | 21.46/0.85 | 17.54/0.72 | 18.73/0.81 |
| **DrTMO [7]** | 62.46 | 56.68 | 59.51 | 21.97/0.84 | 19.96/0.77 | 22.07/0.85 |
| **Deep Recur [12]** | 62.94 | 57.30 | 59.21 | 22.37/0.82 | 20.51/0.76 | 21.66/0.82 |
| **Deep Mask [28] *** | 64.01 | 55.30 | 58.40 | 21.10/0.84 | 17.23/0.73 | 18.73/0.81 |
| **Deep Single [17]** | 64.28 | 57.28 | 59.55 | 24.12/0.86 | 20.48/0.79 | 22.13/0.86 |
| **Deep Synth [10]** | 63.71 | 57.12 | 59.28 | 22.92/0.83 | 20.17/0.77 | 21.67/0.83 |
| **Deep HDRUNet [3]** | 63.87 | 57.24 | 58.96 | 23.17/0.84 | 20.23/0.78 | 21.42/0.81 |
| **Proposed** | **65.20** | **58.92** | **60.48** | **25.17/0.88** | **21.81/0.82** | **22.48/0.87** |

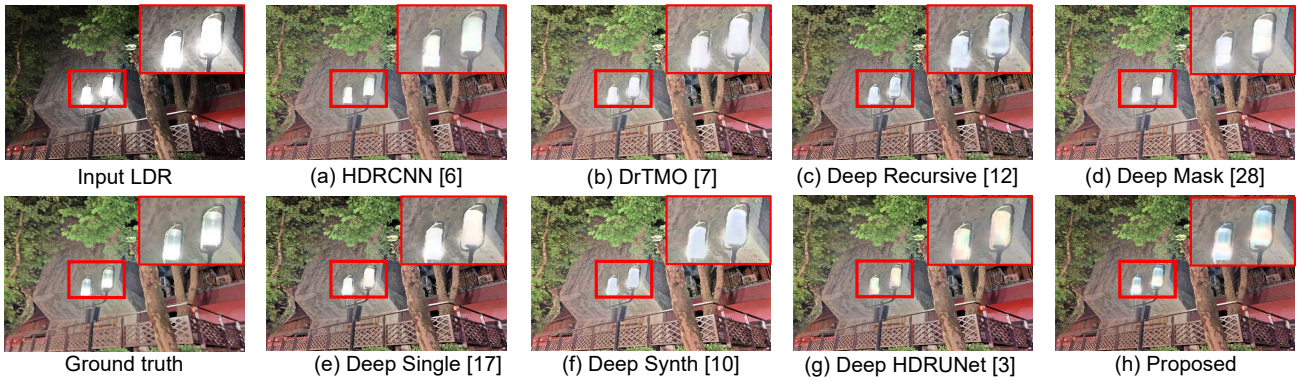


Figure 5. Visual comparison on the dark night scene. The predicted HDR images are tone mapped by [15] for LDR display.

exposed areas. Furthermore, the comparisons between the running time, model parameters, and more visual results can be found in the supplementary materials.

### 4.3. Ablation studies

We evaluate the contributions of individual components and the quantitative experimental results on the HDR-TEST dataset are shown in Table 2.

**Exposure adaptive block.** As shown in Fig. 7, due to the lack of adaptive capacity, the down-exposure model fails to reconstruct uniform and natural luminance and leads to severe artifacts. With the help of the EAB, the decoded features can be modified and thus avoid these artifacts.

**Cross-model attention block.** As Fig. 8 shows, restoring the lost textures in the over/under-exposed regions is a hard task. With the help of the proposed CMAB, the encoder can utilize useful features from the current model and the opposite model, and restore more realistic details. Fig. 9 shows the spatial attention masks calculated by the CMAB of the ADEM, where the $M_{cur}$ focuses on the over-exposed regions to extract the useful features and $M_{ref}$ pays attention to the low-light regions to provide available information from the AUEM. We also remove the $F_{ref}$ from the opposite model to validate the role of the cross-model features, and the HDR-VDP-2.2 result is: w/o $F_{ref}$ (64.52) and with $F_{ref}$ (64.97).

Table 2. Quantitative ablation study of each individual component.

| **Baseline** | **EAB** | **LGC** | **CMAB** | **MEFM** | **HDR-VDP-2** |
|---|---|---|---|---|---|
| ✓ | - | - | - | - | 62.35 |
| ✓ | ✓ | - | - | - | 63.89 |
| ✓ | ✓ | ✓ | - | - | 64.14 |
| ✓ | ✓ | ✓ | ✓ | - | 64.97 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **65.20** |

Table 3. Quantitative ablation study of each training loss.

| **L1** | **Perceptual** | **Progressive** | **MAGAN** |
|---|---|---|---|
| 63.37 | 64.14 | 64.69 | **65.20** |

**Multi-exposure fusion model.** As shown in Table 2, with the proposed MEFM, the defects caused by estimating only two images can be compensated with more precise results. Meanwhile, we also compare the proposed MEFM with the non-learning-based MEF method [4] [10]. Fig. 10 shows the tone-mapped HDR images fused by the pro-

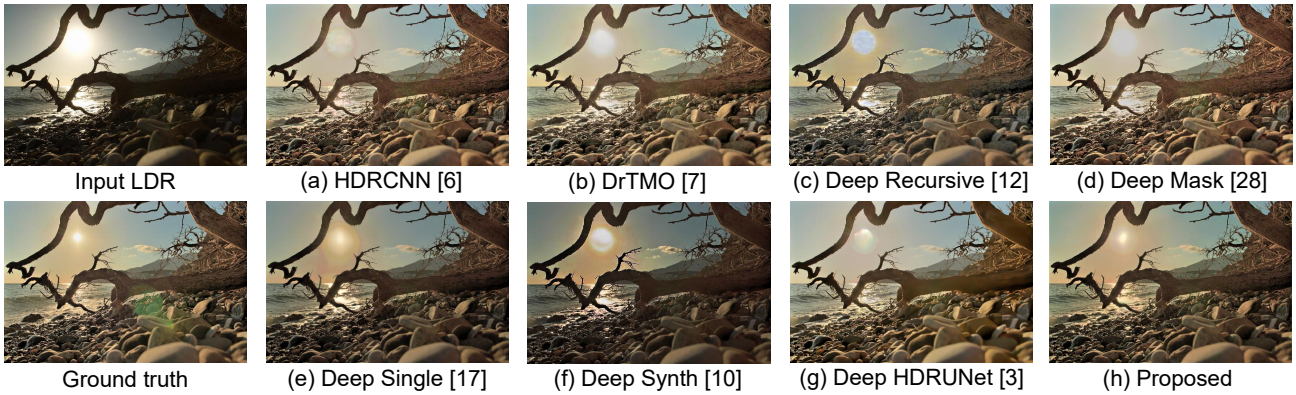


Figure 6. Visual comparison on the daytime scene.

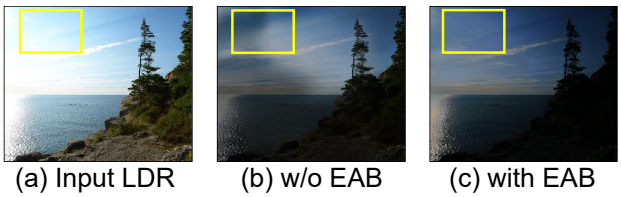


Figure 7. The ablation study of the exposure adaptive block (exposure-down image).

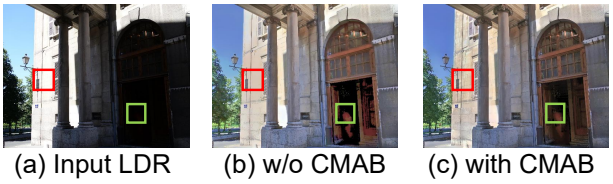


Figure 8. The ablation study of the cross-model attention block.

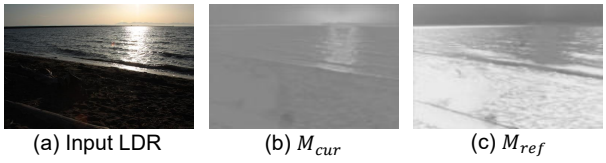


Figure 9. Visual results of the CMAB attention masks.

posed method and [10]. We also compare the running time for merging three multi-exposure images on Tesla V100: MEFM (5ms) and Debevec et al. [4] [10] (1807ms), which demonstrates that the proposed MEFM can generate more accurate results with much less time-complexity.

**Training losses.** We evaluate the contributions of each training loss and Table 3 shows the quantitative results on the HDR-TEST dataset, where MAGAN denotes the mask-aware generative adversarial loss. Due to the page limit, more visual comparisons on the ablation study of training losses can be found in the supplementary materials.

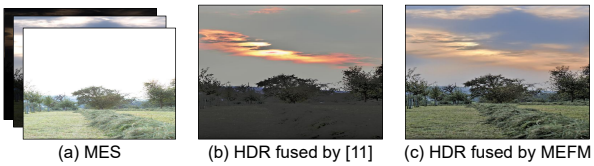


Figure 10. Visual comparison on the multi-exposure fusion results.

## 5. Conclusions

In this paper, we propose a novel inverse tone mapping method that only needs to estimate two exposure images, i.e., one exposure-up and one exposure-down, solving the problem that the optimal length of the multi-exposure stack is difficult to be determined. At first, we propose and open the SICE-S dataset, which can provide the optimal exposure adjustment labels for future stack-based ITM works. Secondly, we design the exposure adaptive block which makes the decoder generate desired results based on the different luminance distributions of the inputs. Thirdly, we devise the cross-model attention block to utilize the information from both of the exposure adjustment models. Finally, we design the learning-based multi-exposure fusion model to produce more accurate HDR radiance fast. Experimental results show that the proposed ITM method can outperform the state-of-the-art ITM methods in both quantitative and qualitative evaluations.

**Acknowledgment** This work is financially supported by National Natural Science Foundation of China U21B2012, 62072013 and 61972129, Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents RCJC20200714114435057, Shenzhen Science and Technology Program-Shenzhen Hong Kong joint funding project of SGDX20211123144400001, This work is also financially supported for Outstanding Talents Training Fund in Shenzhen.

## References


[1] Ronan Boitard, Mahsa T Pourazad, Panos Nasiopoulos, and Jim Slevinsky. Demystifying high-dynamic-range technology: A new evolution in digital media. *IEEE Consumer Electronics Magazine*, 4(4):72–86, 2015. 1

[2] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018. 2, 7

[3] Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong. Hdrunet: Single image hdr reconstruction with denoising and dequantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 354–363, June 2021. 3, 6, 7

[4] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. 2008. 1, 2, 3, 4, 7, 8

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[6] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017. 1, 2, 5, 6, 7

[7] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6):177–1, 2017. 1, 2, 3, 6, 7

[8] Mark D Fairchild. The hdr photographic survey. In *Color and imaging conference*, volume 2007, pages 233–238. Society for Imaging Science and Technology, 2007. 6, 7

[9] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021. 3

[10] Jung Hee Kim, Siyeong Lee, and Suk-Ju Kang. End-to-end differentiable learning to hdr image synthesis for multi-exposure images. *arXiv preprint arXiv:2006.15833*, 2020. 1, 2, 3, 5, 6, 7, 8

[11] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep chain hdri: Reconstructing a high dynamic range image from a single low dynamic range image. *IEEE Access*, 6:49913–49924, 2018. 1, 2, 3

[12] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 596–611, 2018. 1, 2, 3, 5, 6, 7

[13] Siyeong Lee, So Yeon Jo, Gwon Hwan An, and Suk-Ju Kang. Learning to generate multi-exposure stacks with cycle consistency for high dynamic range imaging. *IEEE Transactions on Multimedia*, 2020. 2

[14] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 4

[15] Zhetong Liang, Jun Xu, David Zhang, Zisheng Cao, and Lei Zhang. A hybrid l1-l0 layer decomposition model for tone mapping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4758–4766, 2018. 1, 6, 7

[16] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 3

[17] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1651–1660, 2020. 2, 5, 6, 7

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4

[19] Kede Ma, Hui Li, Hongwei Yong, Zhou Wang, Deyu Meng, and Lei Zhang. Robust multi-exposure image fusion: a structural patch decomposition approach. *IEEE Transactions on Image Processing*, 26(5):2519–2532, 2017. 1

[20] Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011. 6

[21] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017. 6

[22] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, volume 37, pages 37–49. Wiley Online Library, 2018. 2, 5

[23] Hiromi Nemoto, Pavel Korshunov, Philippe Hanhart, and Touradj Ebrahimi. Visual attention in ldr and hdr images. In *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, number CONF, 2015. 6, 7

[24] Shiyu Ning, Hongteng Xu, Li Song, Rong Xie, and Wenjun Zhang. Learning an inverse tone mapping network with a generative adversarial regularizer. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1383–1387. IEEE, 2018. 5

[25] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016. 6

[26] Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski. *High dy-*

*namic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010. 1

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4, 5

[28] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *ACM Transactions on Graphics (TOG)*, 39(4):80–1, 2020. 2, 6, 7

[29] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8207–8216, 2020. 5

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[31] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3

[32] Han Xu, Jiayi Ma, and Xiao-Ping Zhang. Mef-gan: Multi-exposure image fusion via generative adversarial networks. *IEEE Transactions on Image Processing*, 29:7203–7216, 2020. 1

[33] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019. 3

[34] Zhuoran Zheng, Wenqi Ren, Xiaochun Cao, Tao Wang, and Xiuyi Jia. Ultra-high-definition image hdr reconstruction via collaborative bilateral learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4449–4458, 2021. 2, 3