ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Available on IEEE Xplore®

Research Article

https://doi.org/10.1631/ENG.ITEE.2025.0116

Unsupervised single-image high dynamic range rendering via multi-exposure priors

Han WANG[1], Bolun ZHENG[1], Quan CHEN[2]✉, Qianyu ZHANG[1], Tao ZHANG[1], Jiyong ZHANG[1], Xiang TIAN[3]

[1]*School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China*

[2]*College of Artificial Intelligence, Jiaxing University, Jiaxing 314001, China*

[3]*Institute of Advanced Digital Technology and Instrument, Zhejiang University, Hangzhou 310027, China*

Abstract: Reconstructing high dynamic range (HDR) images from a single low dynamic range (LDR) input requires recovering missing information in highlight-clipped and shadow-distorted regions. Existing methods generally rely on sufficient ground-truth HDR images as supervision signals or multi-exposure LDR sequences to improve quality, limiting their flexibility. To address this, we propose USME-HDR, a framework for single-image HDR reconstruction based on multi-exposure priors, where the HDR reconstruction stage is learned without ground-truth HDR supervision. Specifically, an exposure-adjustment network (EAN) is trained in a supervised manner to map a single LDR image to over/under-exposure pairs. Inspired by the Retinex theory, we further decompose the input into a light map and a light feature, which are fed into the EAN as auxiliary inputs for luminance-aware exposure generation. An exposure time ratio guidance mechanism is further introduced to improve luminance fidelity. Finally, the HDR image is synthesized by fusing the original LDR image with generated multi-exposure images, refined through self-supervised optimization. Experiments demonstrate that during the test phase, USME-HDR reconstructs visually compelling HDR images from only a single LDR input, without requiring real low- or high-exposure images.

Key words: High dynamic range (HDR); HDR reconstruction; Single-image HDR; Unsupervised learning; Multi-exposure prior

1 Introduction

Real-world scenes often exhibit extremely high dynamic ranges, with luminance variations spanning several orders of magnitude. However, conventional digital cameras can only capture a limited dynamic range in a single exposure. This limitation inevitably leads to localized over-exposure or under-exposure in resultant images under extreme illumination conditions. To solve this problem, researchers have focused on high dynamic range (HDR) imaging algorithms, aiming to reconstruct complete brightness information for low dynamic range (LDR) images, thereby improving the visual perception quality.

HDR imaging techniques can be broadly classified into two categories based on input requirements: multi-exposure fusion and single-exposure reconstruction. Multi-exposure fusion methods integrate multiple images of the same scene captured at varying exposure levels to extend the effective dynamic range, enabling simultaneous detail recovery in both highlight (oversaturated) and shadow (undersaturated) regions. These methods fundamentally require strict scene consistency. Specifically, any object movement or camera motion during the capture process will induce inter-frame misalignment, consequently leading to motion-induced artifacts in the reconstructed images. In contrast, single-exposure reconstruction methods directly estimate an HDR image from a single LDR image, offering superior operational flexibility and broader applicability. However, the absence of reference data for extreme illumination regions fundamentally limits their reconstruction fidelity, typically resulting in inferior performance compared to multi-exposure approaches.

Currently, most HDR imaging methods adopt supervised learning paradigms, requiring extensive datasets of high-fidelity ground-truth HDR images for training. The high cost

✉ Quan CHEN, chenquan@alu.hdu.edu.cn
Han WANG, https://orcid.org/0009-0001-7042-236X
Bolun ZHENG, https://orcid.org/0000-0001-8788-1725
Quan CHEN, https://orcid.org/0000-0003-2858-6771
Qianyu ZHANG, https://orcid.org/0009-0003-2388-9391
Tao ZHANG, https://orcid.org/0000-0002-7358-0603
Jiyong ZHANG, https://orcid.org/0000-0001-9600-8477
Xiang TIAN, https://orcid.org/0000-0003-0735-8454

CLC number: TP391.4

Received: Nov. 4, 2025; Revision accepted: Apr. 19, 2026;
Crosschecked: Apr. 25, 2026; Published online: May 19, 2026

of acquiring real HDR images makes it difficult for such methods to improve their performance by continuously expanding the dataset. To mitigate dependence on labeled data, unsupervised learning strategies have been explored for HDR reconstruction (Prabhakar et al., 2021; Huang et al., 2022; Yan et al., 2023a; Nazarczuk et al., 2024). However, these methods typically yield synthesized images with suboptimal perceptual quality. Zhang et al. (2024) subsequently proposed SelfHDR, a self-supervised framework that generates HDR images from three LDR inputs with varying exposures, achieving visual fidelity approaching supervised methods. Despite its effectiveness, SelfHDR requires three LDR images during the inference phase, which severely limits its application in real-world scenarios. Consequently, there is an urgent need to develop a highly practical unsupervised HDR imaging algorithm that simultaneously satisfies dual objectives: (1) eliminating dependence on ground-truth HDR images during training and (2) generating an HDR image using a single LDR input during inference while maintaining notable visual quality.

In this paper, we aim to reconstruct HDR images directly from a single LDR image without relying on ground-truth HDR supervision. While SelfHDR (Zhang et al., 2024) has demonstrated the feasibility of unsupervised HDR rendering by generating pseudo-HDR images from multi-exposure LDR inputs, this approach remains susceptible to ghosting artifacts in dynamic scenes. We contend that employing generative networks to transform a single image into multiple images with varying exposures can reduce the dependence of generating pseudo-labels on specific input image quantities and exposure configurations, while ensuring that all generated exposures are derived from the same input image and thus maintain consistent spatial structures, thereby avoiding the misalignment issues commonly observed in multi-frame HDR reconstruction. Therefore, we propose USME-HDR, an unsupervised method for HDR image reconstruction with a single LDR input.

Concretely, the USME-HDR encompasses several components: pseudo-label generation, image luminance estimation, multi-exposure image generation, and image fusion. First, USME-HDR employs the pseudo-label generation strategy (Zhang et al., 2024), which fuses three optical-flow-aligned LDR images to generate pseudo-HDR supervision signals, eliminating dependency on real HDR images during training. Optical-flow-aligned high- and low-exposure LDR images serve as auxiliary supervision for learning multi-exposure priors. Subsequently, USME-HDR incorporates two identical exposure-adjustment networks (EANs) that map the input image to high- and low-exposure LDR images in a supervised manner, facilitating the estimation of an expanded color dynamic range. To further enhance brightness and texture details in the synthesized multi-exposure images, USME-HDR estimates the light map and light feature from the input LDR image and feeds them into the EAN as auxiliary inputs. Meanwhile, exposure time is incorporated into the EAN to guide the model in learning accurate global luminance variations. Finally, the input LDR image, in conjunction with the estimated multi-exposure images, is leveraged to generate an HDR image, as illustrated in Fig. 1. Experimental results validate the effectiveness of our unsupervised HDR rendering scheme for single-image input. Visualization results demonstrate that

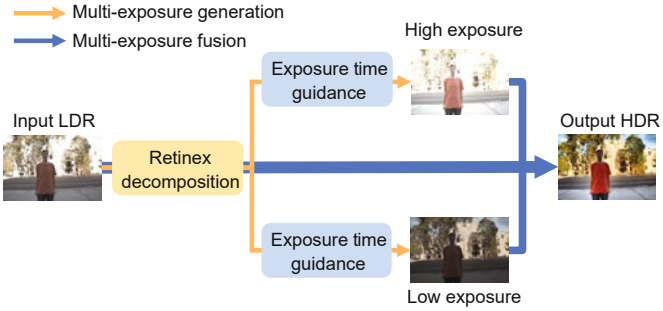

Fig. 1 Overall framework of the proposed USME-HDR network. The network takes a single LDR image as input and estimates high- and low-exposure images through exposure modulation guided by luminance and exposure time. These multi-exposure images are then fused with the input to reconstruct the final HDR image

with a single-image input, our method mitigates the ghosting artifacts associated with multi-exposure image fusion in dynamic scenes. Our contributions are as follows:

1. We pioneer an unsupervised HDR rendering method tailored for single-image input. By learning multi-exposure priors, our method eliminates the dependencies on multi-exposure LDR image sequences while suppressing ghosting artifacts in dynamic scenes.

2. We propose a multi-exposure image generation strategy guided by luminance and exposure time. Specifically, the estimated light map and light feature are incorporated to enhance the brightness and texture details of the synthesized images, while exposure time embeddings guide the model to learn accurate luminance variations.

3. Experimental results demonstrate that the proposed method achieves competitive HDR rendering quality from a single LDR input, rivaling multi-exposure methods. USME-HDR suppresses ghosting artifacts in dynamic scenes, exhibiting significant practical utility.

2 Related works

2.1 Supervised HDR imaging with multi-exposure images

The core challenge in multi-exposure HDR reconstruction is suppressing ghosting artifacts caused by dynamic scene variations. Traditional methods mainly address this issue through misaligned region rejection, image alignment, and patch-based fusion. In recent years, deep neural network (DNN) methods have substantially advanced multi-exposure HDR fusion (Liu SZ et al., 2023; Yan et al., 2023b; Kong et al., 2024; Wu et al., 2024; Chen ZX et al., 2025). Kalantari and Ramamoorthi (2017) introduced the first real-world HDR dataset and proposed a hybrid framework combining optical flow registration and convolutional neural networks (CNNs) for multi-exposure image fusion. Subsequent studies further improved ghosting suppression through spatial attention (Yan et al., 2019) and self-attention mechanisms enabled by Transformers (Liu Z et al., 2022; Tel et al., 2023). For example, Song et al. (2022) adaptively selected Transformer or CNN modules according to regional alignment to improve inference efficiency, while HDR

fusion Transformer (HFT) (Chen RF et al., 2023) adopted a multi-scale hybrid architecture to balance reconstruction performance and computational cost. More recently, generative models have also been introduced into multi-exposure HDR fusion (Yang et al., 2025; Zhu et al., 2025). Hu et al. (2024) proposed a diffusion-based HDR framework that captures low-frequency priors in the latent space and integrates them with dynamic reconstruction networks. Yan et al. (2025a) designed a progressive generation framework combining the segment anything model (SAM) and stable diffusion to synthesize pseudo-static LDR images, followed by dedicated refinement modules for detail enhancement. However, both CNN- and Transformer-based methods fundamentally rely on multi-exposure LDR inputs paired with ground-truth HDR supervision, resulting in resource-intensive data acquisition that limits practical deployment.

2.2 Supervised HDR imaging with single-exposure images

Single-exposure HDR reconstruction aims to generate HDR images from standard dynamic range (SDR) inputs by predicting missing details in over/under-exposed regions (Zheng et al., 2022a, 2022b; Chen SK et al., 2023; Le et al., 2023; Zou et al., 2023; Dille et al., 2024; Xu et al., 2024). A prevalent approach involves inverse tone mapping, synthesizing multi-exposure stacks from single inputs. Endo et al. (2017) employed a dual-branch network to generate high/low-exposure images from middle-exposure LDR inputs. Lee et al. (2018) extended this concept with a cascaded architecture, proposing a relative exposure-guided dual-network framework. However, these methods suffer from limited exposure controllability and inadequate imaging modeling. Alternative strategies leverage UNet architectures or conditional generative networks for direct saturated region recovery via content-aware enhancement. For instance, Santos et al. (2020) proposed a feature masking strategy that mitigates training ambiguity caused by saturation. Liu YL et al. (2020) introduced a multi-stage framework to progressively correct quantization errors and recover highlights, while Khan et al. (2019) adopted a recursive design that expands receptive fields at the cost of high computational complexity. HDRUNet (Chen XY et al., 2021) incorporated spatial feature transformation modules for adaptive detail restoration across varying inputs, enhancing reconstruction fidelity. These methods still rely on real HDR images for model optimization, which incurs expensive data acquisition costs.

2.3 Few-shot and self-supervised HDR reconstruction

To reduce reliance on real HDR data, researchers have explored few-shot and self-supervised strategies for HDR reconstruction. Prabhakar et al. (2021) proposed a method for generating pseudo-HDR-LDR pairs from unlabeled data through HDR prediction and exposure degradation. Nazarczuk et al. (2024) constructed approximate HDR images by fusing well-exposed LDR patches. SelfHDR (Zhang et al., 2024) achieved high-quality HDR generation without ground-truth supervision by aligning multi-exposure images to create pseudo-HDR supervision, with additional color and structural constraints optimizing model performance. Despite mitigating the need for real HDR images, these methods still require multi-exposure inputs during inference, limiting practical deployment flexibility. To address this issue, we propose an unsupervised HDR rendering method for the single-image input. During the test phase, our method synthesizes HDR images from a single LDR image while eliminating dependencies on both real HDR images and multi-exposure LDR images.

3 Proposed method

3.1 Motivation

Current supervised HDR reconstruction methods require real HDR images for training, while unsupervised HDR reconstruction methods require multi-exposure images as input and synthesize pseudo-labels for model optimization. Both paradigms impose data constraints that increase training costs and limit application flexibility. To overcome these limitations, we focus on developing an unsupervised method for high-quality HDR rendering from a single LDR image. This method aims to reduce data acquisition costs, enhance algorithmic versatility, and inherently mitigate ghosting artifacts associated with multi-exposure fusion. Two core challenges must be addressed: (1) Under the unsupervised paradigm, how to generate suitable pseudo-labels for model training? (2) How to estimate rich image details across a wide dynamic range from a single LDR image? A straightforward solution is to leverage the network to learn multi-exposure priors, mapping the input image into LDR images with varying exposure levels, and then to use these multi-exposure LDR images to synthesize HDR images without requiring ground-truth HDR supervision. During the test phase, the model only requires a single LDR image as input and, equipped with learned exposure priors, adjusts its exposure levels to generate a virtual multi-exposure stack, enabling HDR rendering. By learning these priors, the model circumvents the need for complex multi-exposure inputs and enhances the synthesized HDR image's visual quality by estimating richer luminance and texture details. Accordingly, we propose the USME-HDR network. As shown in Fig. 2, it comprises four core components: (1) Pseudo-label generation synthesizes pseudo-HDR images to supervise model optimization; (2) Luminance estimation decomposes the input LDR image into a light map and a light feature to guide subsequent multi-exposure image generation; (3) Multi-exposure image generation generates corresponding high- and low-exposure LDR images, forming a multi-exposure image set; (4) Image fusion synthesizes the HDR image from the generated multi-exposure LDR images. The following sections detail each component.

3.2 Pseudo-label generation

To provide supervision signals for the model, we construct pseudo-HDR images from LDR inputs. We adopt SelfHDR's pseudo-label generation framework (Zhang et al., 2024). Specifically, we denote the LDR image captured with exposure time t_i as I_i, where $i = 1, 2, 3$ and $t_1 < t_2 < t_3$. These LDR

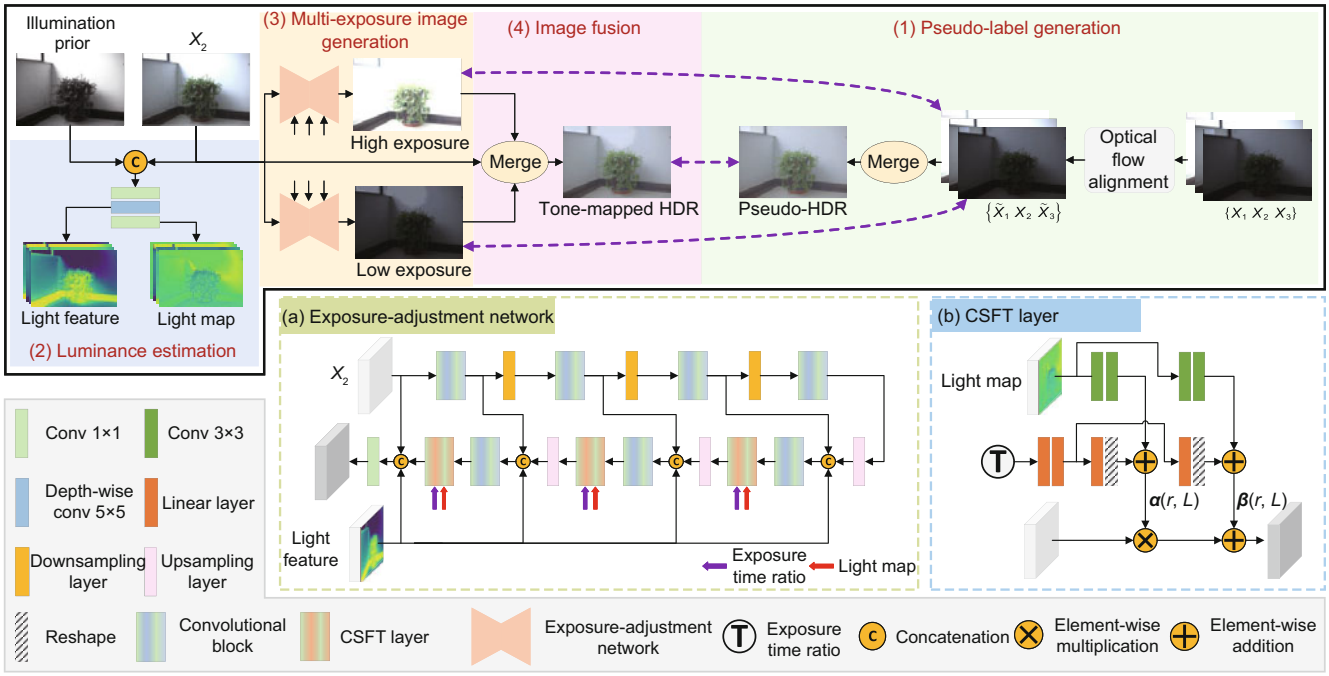

Fig. 2 Overview of the USME-HDR. The input X_2 is decomposed into a light map and a light feature, which are used to guide the subsequent multi-exposure generation process. (a) Architecture of the EAN. The network takes the input image together with luminance-related representations to generate low- and high-exposure images under different exposure levels. The light feature is concatenated with intermediate features at each decoder block, while the light map and exposure time ratio are jointly used as modulation conditions. (b) Structure of the CSFT layer. The light map and exposure time ratio are used to predict the spatially adaptive modulation parameters, which are applied to intermediate features through an affine transformation for luminance-aware exposure modulation. $\tilde{X}_1$ and $\tilde{X}_3$ denote X_1 and X_3 after being aligned with reference to X_2

images are mapped to the HDR domain, defined by

$$H_i = \frac{I_i^{\gamma}}{t_i}, \tag{1}$$

where H_i denotes the HDR-domain radiance image converted from the LDR image I_i, and γ denotes the gamma correction parameter and is generally set to 2.2.

For more general dynamic scenes, an optical flow estimation method is utilized to align multi-exposure images. We take the middle-exposure image I_2 as the reference to estimate optical flows toward I_1 and I_3. Based on the obtained flows, H_1 and H_3 are backward warped to generate the aligned HDR-domain images $\tilde{H}_1$ and $\tilde{H}_3$, , which are approximately aligned with H_2. The pseudo-HDR image I_{PGT} can be generated as follows:

$$I_{\mathrm{PGT}} = \frac{A_1\tilde{H}_1 + A_2H_2 + A_3\tilde{H}_3}{A_1 + A_2 + A_3}, \tag{2}$$

where A_i denotes the pixel-wise fusion weight. Following the approach of Kalantari and Ramamoorthi (2017), A_i is defined as

$$A_1 = 1 - \Lambda_1(I_2), \quad A_2 = \Lambda_2(I_2), \quad A_3 = 1 - \Lambda_3(I_2), \tag{3}$$

where $\Lambda_i(I_2)$ is shown in Fig. 3.

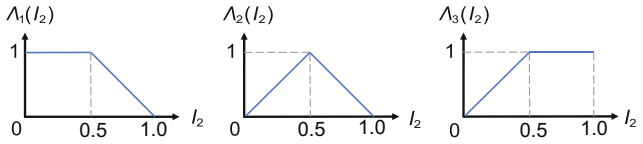

Fig. 3 The triangle functions that are used as the blending weights to generate pseudo-HDR images

3.3 Luminance estimation

We argue that generating multi-exposure images from a single input can be approximately interpreted as luminance adjustment, and luminance information provides important guidance for synthesizing images with appropriate exposure levels. Inspired by the Retinex theory (Land and McCann, 1971), we estimate illumination-related representations from the input image. Specifically, the 6-channel image X_2, formed by concatenating I_2 and H_2, is taken as the input. A single-channel coarse luminance map is first obtained via a mean operation across channels. This map is then concatenated with X_2 to form a 7-channel feature map. The resulting feature map is processed by three convolutional layers. First, a 1×1 convolution is applied for feature compression and channel fusion. Second, a 5×5 depth-wise convolution captures local luminance variations, and its output is defined as the light feature for subsequent feature fusion. Finally, a 1×1 convolution generates a 6-channel light map to provide fine-grained luminance guidance. The light feature and light map are then fed into the EAN to guide the prediction of high- and low-exposure images. This design enables the network to implicitly model the spatial illumination distribution and perform spatially adaptive exposure adjustment without requiring explicit detection of extreme illumination regions.

3.4 Multi-exposure image generation

Single LDR images fail to simultaneously capture complete information in over- and under-exposed regions. To achieve visually plausible luminance distribution and detail restoration in HDR reconstruction, we propose an EAN that learns multi-exposure priors to map a single LDR input to high/low-exposure LDR pairs. It should be noted that the proposed framework is unsupervised with respect to HDR reconstruction, while the multi-exposure generation stage is trained with supervision from aligned real low- and high-exposure images. At inference time, EAN takes only the real middle-exposure image as input, without requiring any real low- or high-exposure images, and predicts the corresponding low- and high-exposure images. As depicted in Fig. 2a, EAN adopts a U-shaped encoder–decoder architecture. Its fundamental building block consists of two stacked 3×3 convolutional layers (with Mish activation), while downsampling is implemented via max-pooling and upsampling via PixelShuffle layers. To enhance the network's perception of illumination distribution, we inject the light feature into the decoder. This design mitigates quality degradation in the generated images caused by supervision signal misalignment through reinforced luminance awareness. To enable adaptive feature luminance modulation conditioned on exposure regions, we design a composite spatial feature transform (CSFT) layer within the decoder stage, whose structure is detailed in Fig. 2b. The exposure time ratio is defined as

$$r_{\mathrm{up}} = \frac{t_3}{t_2},\ r_{\mathrm{down}} = \frac{t_1}{t_2},\ t_1 < t_2 < t_3, \tag{4}$$

where r_{up} and r_{down} denote the exposure time ratios of the high- and low-exposure images relative to the middle-exposure image, respectively. Subsequently, this signal is encoded into a high-dimensional vector via two fully-connected layers. The light map is also utilized as a spatial modulation:

$$\boldsymbol{F}' = \boldsymbol{\alpha}(r, L) \otimes \boldsymbol{F} + \boldsymbol{\beta}(r, L), \tag{5}$$

where $\otimes$ denotes the element-wise multiplication. $\boldsymbol{F}' \in \mathbb{R}^{H\times W\times C}$ represents the modulated feature map and $\boldsymbol{F}$ denotes the input feature map before modulation. $\boldsymbol{\alpha}(r, L) \in \mathbb{R}^{H\times W\times C}$ and $\boldsymbol{\beta}(r, L) \in \mathbb{R}^{H\times W\times C}$ denote the scale and shift terms computed based on the exposure time ratio and the light map, respectively. r is the exposure time ratio, L is the light map, and H, W, and C denote the height, width, and number of channels of the feature map, respectively.

To ensure that the generated high- and low-exposure images closely match the real exposure levels, we supervise the EAN using the LDR images $\tilde{I}_1$ and $\tilde{I}_3$ aligned to I_2 via optical flow. The corresponding supervision losses, defined as the low- and high-exposure reconstruction losses $\mathcal{L}_{\mathrm{le}}$ and $\mathcal{L}_{\mathrm{he}}$, are given by

$$\mathcal{L}_{\mathrm{le}}(I_2^{\mathrm{low}}, \tilde{I}_1) = \left\| \left(\mathcal{T}(I_2^{\mathrm{low}}) - \mathcal{T}(\tilde{I}_1) \right) \otimes M_{\mathrm{lem}} \right\|_1, \tag{6}$$

$$\mathcal{L}_{\mathrm{he}}(I_2^{\mathrm{high}}, \tilde{I}_3) = \left\| \left(\mathcal{T}(I_2^{\mathrm{high}}) - \mathcal{T}(\tilde{I}_3) \right) \otimes M_{\mathrm{hem}} \right\|_1, \tag{7}$$

where $\|\cdot\|_1$ denotes the L_1 norm for computing the pixel-wise absolute difference. $\mathcal{T}(\cdot)$ denotes a tone-mapping function. Given an HDR image Y, it is defined as

$$\mathcal{T}(Y) = \frac{\ln(1+\mu Y)}{\ln(1+\mu)},\ \mu = 5000. \tag{8}$$

I_2^{low} and I_2^{high} denote the EAN outputs for the low- and high-exposure reconstruction, respectively. M_{lem} and M_{hem} are binary masks that select well-aligned pixels in $\tilde{I}_1$ and $\tilde{I}_3$ to prevent incorrect supervision, respectively. Each pixel in the masks is defined as

$$M_{\mathrm{lem}}^p = \begin{cases} 1, & \left| \left(\left(\mathcal{T}(\tilde{H}_1) - \mathcal{T}(H_2) \right) \otimes \Lambda_2(I_2) \right)^p \right| < \sigma_{\mathrm{le}}, \\ 0, & \left| \left(\left(\mathcal{T}(\tilde{H}_1) - \mathcal{T}(H_2) \right) \otimes \Lambda_2(I_2) \right)^p \right| \geq \sigma_{\mathrm{le}}, \end{cases} \tag{9}$$

$$M_{\mathrm{hem}}^p = \begin{cases} 1, & \left| \left(\left(\mathcal{T}(\tilde{H}_3) - \mathcal{T}(H_2) \right) \otimes \Lambda_2(I_2) \right)^p \right| < \sigma_{\mathrm{he}}, \\ 0, & \left| \left(\left(\mathcal{T}(\tilde{H}_3) - \mathcal{T}(H_2) \right) \otimes \Lambda_2(I_2) \right)^p \right| \geq \sigma_{\mathrm{he}}, \end{cases} \tag{10}$$

where p denotes an individual pixel location in the mask. Both σ_{le} and σ_{he} are thresholds and are set to $5/255$.

3.5 Image fusion

For multi-exposure image fusion, we also employ the exposure-weighted fusion method. Prior study (Debevec and Malik, 1997) has demonstrated that this method can effectively integrate information from both bright and dark regions of input LDR images, thereby reconstructing HDR images with high visual quality and wide dynamic range. Given its effectiveness and simplicity, we directly apply this fusion strategy without introducing an additional network. The final fused HDR image is defined as $\hat{Y}$.

Following SelfHDR (Zhang et al., 2024), we leverage both the pseudo-HDR image and the middle-exposure image to guide network learning. The middle-exposure image preserves structural details via a structure-preserving loss $\mathcal{L}_{\mathrm{sp}}$, defined as

$$\mathcal{L}_{\mathrm{sp}}(\hat{Y}, H_2) = \left\| \left(\mathcal{T}(\hat{Y}) - \mathcal{T}(H_2) \right) \otimes M_{\mathrm{sp}} \right\|_1, \tag{11}$$

where $M_{\mathrm{sp}} = \Lambda_2(I_2)$ (see Fig. 3), which emphasizes well-exposed regions to mitigate the impact of under- and over-exposed areas in the reference H_2. Meanwhile, the image I_{PGT} encourages structural learning from non-reference inputs through a structure-expansion loss $\mathcal{L}_{\mathrm{se}}$, defined as

$$\mathcal{L}_{\mathrm{se}}(\hat{Y}, I_{\mathrm{PGT}}) = \left\| \left(\mathcal{T}(\hat{Y}) - \mathcal{T}(I_{\mathrm{PGT}}) \right) \otimes M_{\mathrm{se}} \right\|_1, \tag{12}$$

where M_{se} denotes a binary mask indicating whether each pixel in I_{PGT} is formed from well-aligned multi-exposure images. Adhering to SelfHDR, each pixel M_{se}^p of M_{se} is defined as

$$M_{\mathrm{se}}^p = \begin{cases} 1, & |((\mathcal{T}(I_{\mathrm{PGT}}) - \mathcal{T}(H_2)) \otimes \Lambda_2(I_2))^p| < \sigma_{\mathrm{se}}, \\ 0, & |((\mathcal{T}(I_{\mathrm{PGT}}) - \mathcal{T}(H_2)) \otimes \Lambda_2(I_2))^p| \geq \sigma_{\mathrm{se}}, \end{cases} \tag{13}$$

where σ_{se} denotes a threshold and is set to $5/255$.

To reduce the impact of saturation during training, we adopt an exposure-aware loss $\mathcal{L}_{\mathrm{ea}}$ that adaptively weights supervision based on pixel brightness, defined as

$$\mathcal{L}_{\mathrm{ea}}(\hat{Y}, I_{\mathrm{PGT}}) = \left\| \left(\mathcal{T}(\hat{Y}) - \mathcal{T}(I_{\mathrm{PGT}}) \right) \otimes M_{\mathrm{ep}} \right\|_1, \tag{14}$$

where M_{ep} denotes an exposure-aware weight mask applied to the reference middle-exposure image (Santos et al., 2020). Pixels are categorized based on their brightness E, and assigned weights as follows:

$$M_{\mathrm{ep}} = \begin{cases} 1.0, & \text{if } 0.1 \leq E \leq 0.9, \\ 0.1, & \text{if } 0.05 \leq E < 0.1 \text{ or } 0.9 < E \leq 0.95, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Finally, a visual geometry group (VGG)-based perceptual loss $\mathcal{L}_{\mathrm{p}}$ (Simonyan and Zisserman, 2015) is used to enhance the perceptual quality of the reconstructed HDR image, defined as

$$\mathcal{L}_{\mathrm{p}}(\hat{Y}, I_{\mathrm{PGT}}) = \sum_{k\in\mathcal{K}} \left\| \phi_k(\mathcal{T}(\hat{Y})) - \phi_k(\mathcal{T}(I_{\mathrm{PGT}})) \right\|_1, \quad (16)$$

where $\phi_k(\cdot)$ denotes the output of the k^{th} layer in the VGG network, and $\mathcal{K}$ denotes the set of the selected VGG feature layers used for computing the perceptual loss. In short, the reconstruction network parameters Θ_{R} are optimized as

$$\Theta_{\mathrm{R}}^{*} = \arg\min_{\Theta_{\mathrm{R}}} \left(\lambda_{\mathrm{le}}\mathcal{L}_{\mathrm{le}} + \lambda_{\mathrm{he}}\mathcal{L}_{\mathrm{he}} + \lambda_{\mathrm{sp}}\mathcal{L}_{\mathrm{sp}} + \mathcal{L}_{\mathrm{se}} + \mathcal{L}_{\mathrm{ea}} + \mathcal{L}_{\mathrm{p}} \right), \quad (17)$$

where Θ_{R}^{*} denotes the optimized parameters obtained by minimizing the objective function. λ_{sp} denotes the weight coefficient of the structure-preserving loss and is set to 4, while λ_{le} and λ_{he} denote the weight coefficients for the low- and high-exposure reconstruction losses, respectively, and are both set to 2.

4 Experiments

4.1 Implementation details

4.1.1 Training details

During training, image patches of size 128×128 are randomly cropped from the original images as input. The batch size is fixed at 16, and the model is trained for 150 epochs using the Adam optimizer. The initial learning rate is set to 1×10^{-4}, decaying by half every 50 epochs. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

4.1.2 Datasets

The experiments are conducted on the RealHDRV dataset (Shu et al., 2024) and Kalantari's dataset (Kalantari and Ramamoorthi, 2017).

1. RealHDRV dataset. It consists of 450 training samples and 50 test samples, covering a wide range of scenes including daytime, nighttime, indoor, and outdoor environments. Each sample contains three LDR images captured at different exposure values, either {−2EV, 0EV, +2EV} or {−3EV, 0EV, +3EV} (here, EV stands for exposure value, a measure of relative exposure), along with a corresponding high-quality HDR ground truth.

2. Kalantari's dataset. It includes 74 training and 15 test samples of dynamic scenes. Each sample provides three differently exposed LDR images and the corresponding ground-truth HDR image.

3. Our real-world test dataset. To further evaluate the robustness of the proposed method in real-world scenarios, we collect a real-world test dataset using a Canon 200D II camera. The dataset contains 40 groups of LDR image samples, covering both indoor and outdoor scenes with diverse illumination conditions and dynamic range variations. This dataset is used only for testing real-world performance. Representative examples are shown in Fig. 4.

We conduct training and evaluation on the public datasets to assess the performance of the proposed method, and further test it on our real-world dataset to examine its robustness under practical imaging conditions.

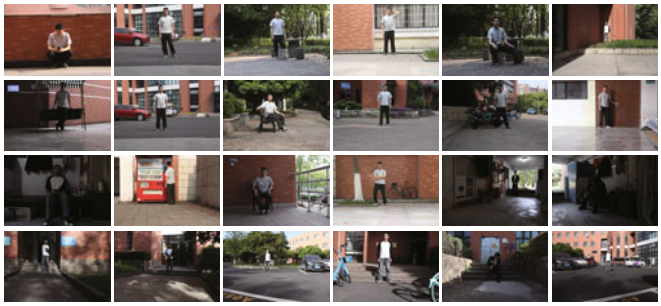

Fig. 4 Representative examples from our self-captured real-world test dataset

4.1.3 Evaluation configurations

We adopt peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) as evaluation metrics. Results computed in the linear and tone-mapped domains are distinguished by "-L" and "-μ," respectively. Furthermore, we employ HDR-VDP-2 (Mantiuk et al., 2011) to quantify the perceptual difference between the output and the ground truth, where a higher HDR-VDP-2 score indicates better perceptual quality. The metric is computed using the official implementation with an sRGB display model, assuming a 24-inch display and a 0.5-m viewing distance. For real-world images without HDR ground truth, we further adopt four no-reference image quality metrics, including multi-scale image quality Transformer (MUSIQ) (Ke et al., 2021), contrastive language image pre-training (CLIP)-based image quality assessment (CLIP-IQA) (Wang JY et al., 2023), natural image quality evaluator (NIQE) (Mittal et al., 2013), and blind/referenceless image spatial quality evaluator (BRISQUE) (Mittal et al., 2012), to evaluate perceptual quality. For MUSIQ and CLIPIQA, higher values indicate better image quality, whereas for NIQE and BRISQUE, lower values indicate better perceptual quality.

4.2 Evaluation on the public datasets

We compare the proposed USME-HDR with several representative methods on two public HDR benchmarks, including RealHDRV dataset and Kalantari's dataset. The compared methods include single-image HDR reconstruction methods, i.e., HDRUNet (Chen XY et al., 2021), knowledge-inspired UNet (KUNet) (Wang H et al., 2022), and continuous exposure value representation (CEVR) (Chen SK et al., 2023), as well as representative low-light image enhancement methods, including Retinexformer (Cai et al., 2023), RetinexMamba (Bai et al., 2024), and collaborative transformation framework (CoTF) (Li et al., 2024). For a fair comparison, we consider two experimental settings: (1) direct HDR reconstruction from a single middle-exposure LDR image and (2) indirect reconstruction, where low- and high-exposure images are first generated from the middle-exposure input and the final HDR image is then obtained using the same exposure fusion strategy as ours. For a more comprehensive evaluation, some methods use both 3-channel and 6-channel inputs. The 3-channel input corresponds to the original middle-exposure image I_2, whereas the 6-channel input, denoted as X_2, is formed by concatenating I_2 with its HDR-domain representation along the channel dimension.

The quantitative results on the two public datasets are summarized in Table 1. We first compare USME-HDR with the other representative methods listed in the table. Overall, the proposed method achieves the best or highly competitive performance on both public datasets. In particular, USME-HDR achieves the best overall results on both RealHDRV and Kalantari's datasets under the 6-channel setting. In addition to reconstruction quality, USME-HDR requires fewer multiply-accumulate operations (MACs) and shorter inference time than several competing methods while maintaining a superior quantitative performance, indicating a favorable trade-off between reconstruction quality and computational cost.

The visual comparisons in Fig. 5 further support the quantitative results. On both RealHDRV and Kalantari's datasets, our method better restores structural details and luminance transitions in over-exposed regions, while preserving more natural colors and fewer reconstruction artifacts. These results indicate that the proposed luminance-guided exposure generation and exposure-time-ratio-aware modulation provide more reliable complementary information for HDR reconstruction under challenging illumination conditions.

As SelfHDR is reported under both real multi-exposure and unified-input settings, we discuss its results separately for clarity. To further analyze ghosting robustness in dynamic scenes, we conduct the following experiment under the real multi-exposure setting. Specifically, we compare three HDR reconstruction strategies: (1) SelfHDR, which directly reconstructs HDR images from three real exposure inputs; (2) an optical-flow-aligned exposure fusion baseline, which first aligns the real low- and high-exposure images to the middle-exposure image and then reconstructs HDR using the same exposure-weighted fusion strategy as ours; (3) the proposed USME-HDR, which takes only the middle-exposure image as the real input and generates the corresponding low- and high-exposure images, which are then fused into an HDR image. As shown in Fig. 6 and Table 1, although SelfHDR achieves higher quantitative scores on most evaluation metrics under this setting, methods relying on real multi-exposure inputs (e.g., SelfHDR and optical-flow-aligned fusion) may still suffer from ghosting artifacts in dynamic scenes due to imperfect frame alignment. Since these methods directly exploit three real exposure images, they have access to richer color and exposure information and therefore show stronger color recovery ability. In contrast, the generated exposures in USME-HDR are all derived from the same middle-exposure image, and thus remain spatially consistent with the input, which effectively alleviates ghosting during exposure fusion. These results suggest that, although real multi-exposure methods may achieve better reconstruction fidelity in some cases, our single-frame-driven strategy provides a more robust solution for suppressing artifacts in dynamic scenes.

To provide a more strictly fair comparison, we further evaluate SelfHDR under the unified-input setting. Specifically, we use the middle-exposure image as the only real input, generate the corresponding low- and high-exposure images using our EAN, and then feed these three images into SelfHDR for HDR reconstruction. Under this setting, both methods perform HDR reconstruction using the same set of exposure images generated from the middle-exposure input, thereby eliminating the additional information advantage introduced by real multi-exposure inputs. As shown in Fig. 7 and Table 1, the two methods achieve comparable performance on Kalantari's dataset,

Table 1 Quantitative comparison results on Kalantari's and RealHDRV datasets

Method	Kalantari's dataset					RealHDRV dataset					Computational cost		
	PSNR-μ	PSNR-L	SSIM-μ	SSIM-L	HDR-VDP-2	PSNR-μ	PSNR-L	SSIM-μ	SSIM-L	HDR-VDP-2	N_{MACs} ($\times 10^9$)	N_{params} ($\times 10^6$)	Time (s)
HDRUNet	33.38	32.72	0.9728	0.9455	61.39	29.67	29.18	0.9476	0.9544	60.93	532.70	1.651	0.18
KUNet	30.05	32.42	0.9534	0.9475	61.76	28.97	28.73	0.9214	0.9483	61.44	964.42	1.137	0.215
CoTF	29.64	32.01	0.9723	0.9424	53.99	28.83	28.53	0.9384	0.9558	53.54	2.47	0.31	0.04
Retinexformer	33.93	34.08	0.9758	0.9645	61.64	31.04	29.58	0.9517	0.9508	61.33	389.52	1.606	0.435
RetinexMamba	29.23	33.41	0.9529	0.9418	60.96	31.56	29.73	0.9545	0.9531	62.52	869.35	3.588	1.66
HDRUNet*	36.04	33.63	0.9829	0.9735	60.45	32.32	29.32	0.9727	0.9753	61.48	1065.00	3.190	0.405
KUNet*	34.77	33.49	0.9813	0.9726	59.91	31.04	28.68	0.9735	0.9720	61.22	1929.00	2.273	0.44
CEVR	40.60	36.79	0.9818	0.9747	61.13	37.30	31.32	0.9561	0.9834	61.72	1819.00	6.586	0.329
CoTF*	32.94	33.26	0.9813	0.9640	59.89	34.01	29.14	0.9743	0.9763	60.72	4.97	0.63	0.09
Retinexformer*	36.93	34.20	0.9853	0.9723	61.76	33.85	29.09	0.9714	0.9765	61.04	781.04	3.212	0.92
RetinexMamba*	36.40	33.46	0.9839	0.9709	61.52	34.45	29.91	0.9742	0.9761	61.80	1739.00	7.176	3.21
USME-HDR*	40.92	37.89	0.9872	0.9792	62.47	40.62	32.60	**0.9834**	0.9879	63.07	487.59	3.928	0.16
HDRUNet**	41.12	37.46	0.9871	0.9753	62.47	40.17	31.86	0.9832	0.9851	60.83	1076.00	3.196	0.37
KUNet**	41.18	37.41	0.9873	0.9768	61.99	40.19	32.02	<u>0.9833</u>	0.9854	60.71	1934.00	2.277	0.425
Retinexformer**	41.22	37.46	0.9874	0.9780	61.74	40.52	32.35	0.9829	0.9868	62.70	788.94	3.228	0.98
SelfHDR (multi-exposure)	**43.68**	**41.09**	**0.9901**	**0.9873**	**64.57**	**41.91**	**37.65**	0.9751	**0.9911**	**63.97**	1871.00	1.242	0.353
SelfHDR (unified-input)	<u>42.07</u>	37.50	<u>0.9888</u>	<u>0.9798</u>	61.72	39.68	32.83	0.9748	<u>0.9882</u>	62.77	1871.00	1.242	0.351
USME-HDR**	41.36	<u>38.12</u>	0.9879	0.9794	<u>62.52</u>	<u>40.78</u>	<u>33.01</u>	**0.9834**	0.9882	<u>63.23</u>	494.36	3.932	0.162

Methods marked with $*$ denote indirect reconstruction, where low- and high-exposure images are generated from the middle-exposure input and then fused into an HDR image. Methods marked with $**$ indicate indirect reconstruction with the 6-channel input X_2. Please refer to Section 4.2 for details. For SelfHDR, results under both the original multi-exposure setting and the unified-input setting are reported. The reported time corresponds to processing an image of resolution 1500×1000 on a single RTX 3090 GPU, and N_{MACs} denotes the computational cost under the same configuration. The best and second-best results are highlighted in bold and underlined, respectively

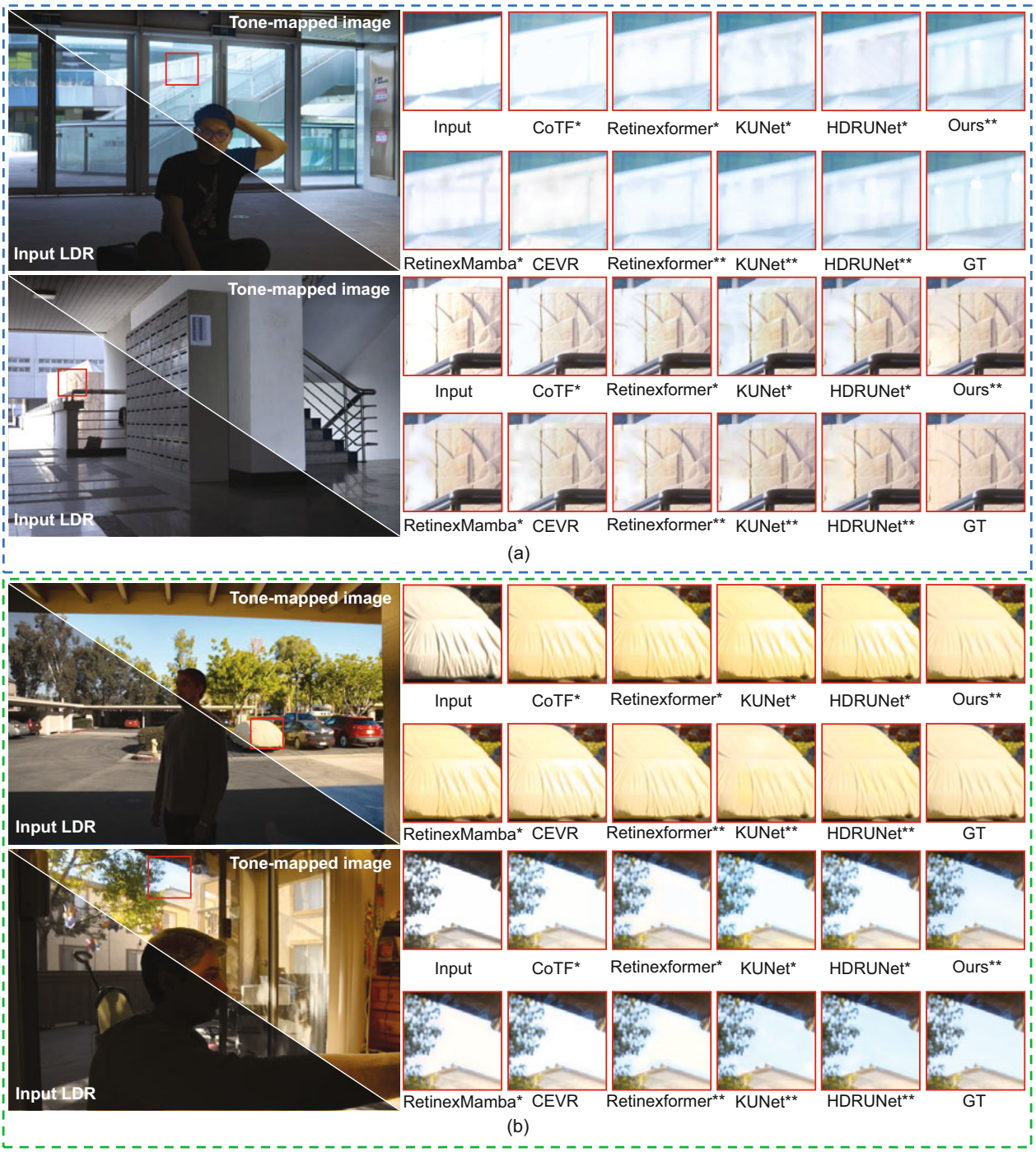

Fig. 5 Visual comparison on the public datasets: (a) RealHDRV dataset; (b) Kalantari's dataset. Compared with other methods, our approach better reconstructs details in over-exposed regions while preserving more natural and consistent colors. GT: ground truth. Methods marked with $*$ denote indirect reconstruction, where low- and high-exposure images are generated from the middle-exposure input and then fused into an HDR image. Methods marked with $**$ indicate indirect reconstruction with the 6-channel input X_2

while our method performs better on PSNR-L. On the RealHDRV dataset, our method achieves better results on PSNR-μ, PSNR-L, and SSIM-μ, while remaining comparable on SSIM-L. This difference may be attributed to the fact that SelfHDR can exploit learned feature priors to generate smoother results in severely saturated regions, whereas our physically constrained fusion is more effective at preserving luminance consistency in regions with valid exposure information.

4.3 Cross-dataset generalization evaluation

To further evaluate the generalization ability of the proposed method under different data distributions, we conduct cross-dataset validation experiments on Kalantari's and RealHDRV datasets. Specifically, the model is trained on one dataset and directly tested on the other without any additional fine-tuning. The quantitative comparison results are reported

Fig. 6 Visual comparison of different HDR reconstruction strategies under real multi-exposure inputs. Yellow dashed boxes denote the results of methods relying on real multi-exposure inputs, while blue dashed boxes denote the results of the proposed USME-HDR. GT: ground truth. SelfHDR$_{\mathbf{AHDRNet}}$ denotes the SelfHDR framework using AHDRNet as the HDR reconstruction network. Methods marked with ∗∗ indicate indirect reconstruction with the 6-channel input X_2

Fig. 7 Visual comparison with SelfHDR$_{\mathbf{AHDRNet}}$ under the unified-input setting. Both methods use the same exposure images generated from the middle-exposure input. GT: ground truth. SelfHDR$_{\mathbf{AHDRNet}}$ denotes the SelfHDR framework using AHDRNet as the HDR reconstruction network. Methods marked with ∗∗ indicate indirect reconstruction with the 6-channel input X_2

Table 2 Quantitative results of cross-dataset validation on Kalantari's and RealHDRV datasets

Method	Train on Kalantari's dataset, test on the RealHDRV dataset				Train on the RealHDRV dataset, test on Kalantari's dataset			
	PSNR-μ	PSNR-L	SSIM-μ	SSIM-L	PSNR-μ	PSNR-L	SSIM-μ	SSIM-L
HDRUNet*	33.90	29.03	0.9671	0.9765	30.28	33.40	0.9777	0.9609
HDRUNet**	<u>37.48</u>	30.59	<u>0.9786</u>	<u>0.9829</u>	39.43	34.92	0.9833	0.9655
KUNet*	32.84	28.48	0.9731	0.9744	29.71	32.89	0.9767	0.9599
KUNet**	37.14	29.70	<u>0.9786</u>	0.9825	<u>39.51</u>	<u>34.96</u>	<u>0.9835</u>	**0.9662**
CEVR	36.70	<u>30.99</u>	0.9519	0.9822	39.48	34.91	0.9815	0.9648
USME-HDR**	**38.11**	**31.21**	**0.9800**	**0.9853**	**39.61**	**35.03**	**0.9852**	<u>0.9658</u>

Methods marked with ∗ denote indirect reconstruction, where low- and high-exposure images are generated from the middle-exposure input and then fused into an HDR image. Methods marked with ∗∗ indicate indirect reconstruction with the 6-channel input X_2. The best and second-best results are highlighted in bold and underlined, respectively

in Table 2. It can be observed that the proposed method achieves the best performance on most evaluation metrics. These results indicate that the proposed method is able to maintain good generalization capability across different data distributions.

4.4 Evaluation on the self-captured real-world test dataset

To further evaluate the practical performance of the proposed method in real-world scenarios, we conduct experiments on our self-captured test dataset. The no-reference image quality evaluation results are reported in Table 3. The proposed method achieves superior or competitive performance across all metrics. The qualitative comparisons in Fig. 8 further support the quantitative results. Compared with other methods, our approach produces more natural luminance transitions and preserves richer details. These observations indicate that the proposed luminance-guided exposure generation strategy remains effective in real complex scenes, validating robustness and practical applicability of our method.

Table 3 Quantitative comparison of no-reference image quality metrics on the self-captured real-world test dataset

Method	MUSIQ(↑)	CLIPIQA(↑)	NIQE(↓)	BRISQUE(↓)
HDRUNet*	60.9387	0.4639	2.7665	20.2884
HDRUNet**	61.6929	0.4882	2.6907	20.2326
KUNet*	61.1535	0.4332	2.6389	19.9659
KUNet**	61.7027	0.4978	2.5858	20.2611
CEVR	60.8931	0.5021	2.7332	18.4003
SelfHDR$_{\text{AHDRNet}}$	61.2132	0.5024	2.5903	18.0960
USME-HDR**	**61.8276**	**0.5081**	**2.5281**	**17.4714**

SelfHDR$_{\text{AHDRNet}}$ denotes the SelfHDR framework using AHDRNet as the HDR reconstruction network. ↑: the higher the better; ↓: the lower the better. Methods marked with ∗ denote indirect reconstruction, where low- and high-exposure images are generated from the middle-exposure input and then fused into an HDR image. Methods marked with ∗∗ indicate indirect reconstruction with the 6-channel input X_2. The best and second-best results are highlighted in bold and underlined, respectively

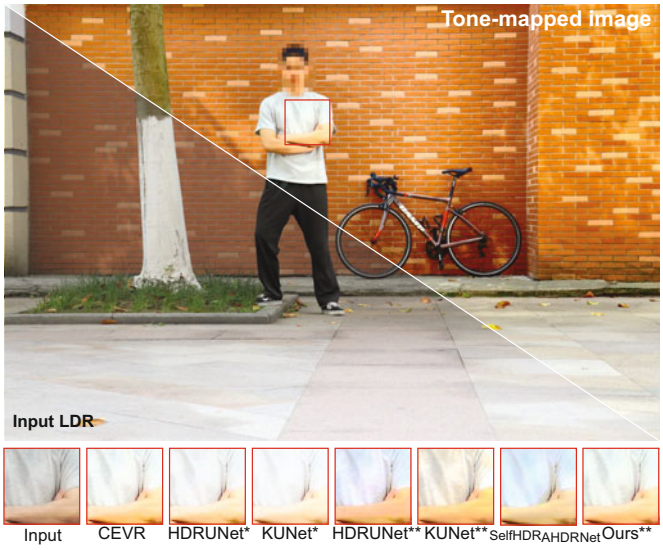

Fig. 8 Visual comparison of our self-captured real-world test dataset. SelfHDR$_{\text{AHDRNet}}$ denotes the SelfHDR framework using AHDRNet as the HDR reconstruction network. Methods marked with ∗ denote indirect reconstruction, where low- and high-exposure images are generated from the middle-exposure input and then fused into an HDR image. Methods marked with ∗∗ indicate indirect reconstruction with the 6-channel input X_2

4.5 Evaluation of the generated multi-exposure images

To further verify the physical plausibility of the generated multi-exposure images, we present a visual comparison between the generated exposures and their corresponding real multi-exposure images in Fig. 9. The results show that in most scenes, the generated high- and low-exposure images are highly consistent with the real exposure images in terms of luminance distribution and structural preservation. These results indicate that the generated multi-exposure images can provide reliable complementary exposure information for the subsequent HDR fusion process.

4.6 Ablation study

All ablation experiments are conducted on the 6-channel input image X_2, unless otherwise specified. We design a series of experiments to evaluate the contributions of different components in our framework.

4.6.1 Effect of luminance estimation

To evaluate the role of the luminance estimation module in the reconstruction process, we design three ablation settings: (1) without (w/o) light map, where the spatial modulation guided by the light map is removed during the decoding stage; (2) w/o light feature, where the light feature is no longer concatenated with the residual features at each decoder level; (3) w/o light map and light feature. As shown in Table 4, the full configuration achieves the best quantitative results across all evaluation metrics, verifying effectiveness of the luminance estimation module in improving HDR reconstruction quality. Fig. 10 further presents the learned light feature and light map produced by the luminance estimation module, together with the exposure generation and final HDR reconstruction results under different ablation settings. For clearer comparison, zoomed-in views of over-exposed regions are also provided. The results show that these luminance-related representations can capture the spatial illumination distribution of the scene and guide the EAN to learn more appropriate region-wise exposure adjustments. The PSNR values computed on the zoomed-in regions further indicate that the full model consistently outperforms all ablated variants. These results verify the effectiveness of the proposed luminance estimation module in both exposure generation and HDR reconstruction.

Table 4 Quantitative results of the effects of the light map, light feature, and exposure time ratio on the RealHDRV dataset

Light map	Light feature	Exposure time ratio	PSNR-μ	PSNR-L	SSIM-μ	SSIM-L
×	×	×	39.32	31.37	0.9817	0.9792
×	×	✓	39.88	32.23	0.9821	0.9864
✓	×	✓	40.44	32.53	0.9827	0.9869
×	✓	✓	40.42	32.69	0.9826	0.9875
✓	✓	✓	**40.78**	**33.01**	**0.9834**	**0.9882**

The best and second-best results are highlighted in bold and underlined, respectively

4.6.2 Effect of different luminance estimation methods

To further investigate whether different luminance estimation strategies influence the overall HDR reconstruction performance, we compare the proposed luminance estimation module with several alternative methods, including: (1) global statistical brightness (GSB), which uses the global mean and standard deviation of the image as luminance statistics; (2) Gaussian-smoothed illumination (GSI), which approximates the illumination distribution using Gaussian-filtered low-frequency components; (3) horizontal/vertical-intensity-based intensity extraction (HVI-I) (Yan et al., 2025b), which uses the intensity channel in the HVI color space as the luminance representation; (4) Retinex decomposition (RetinexNet) (Wei et al., 2018), which estimates reflectance and illumination components through Retinex decomposition.

The quantitative comparison results are reported in Table 5. It can be observed that simple global statistical

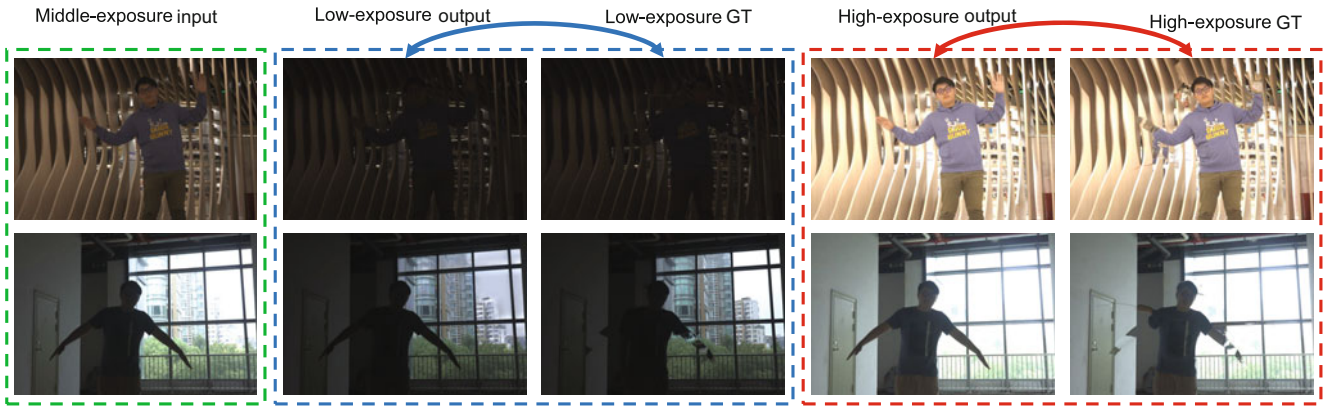

Fig. 9 Visual comparison between the generated and real multi-exposure images. The comparisons for the low-exposure images are marked with a blue dashed box, while those for the high-exposure images are marked with a red dashed box

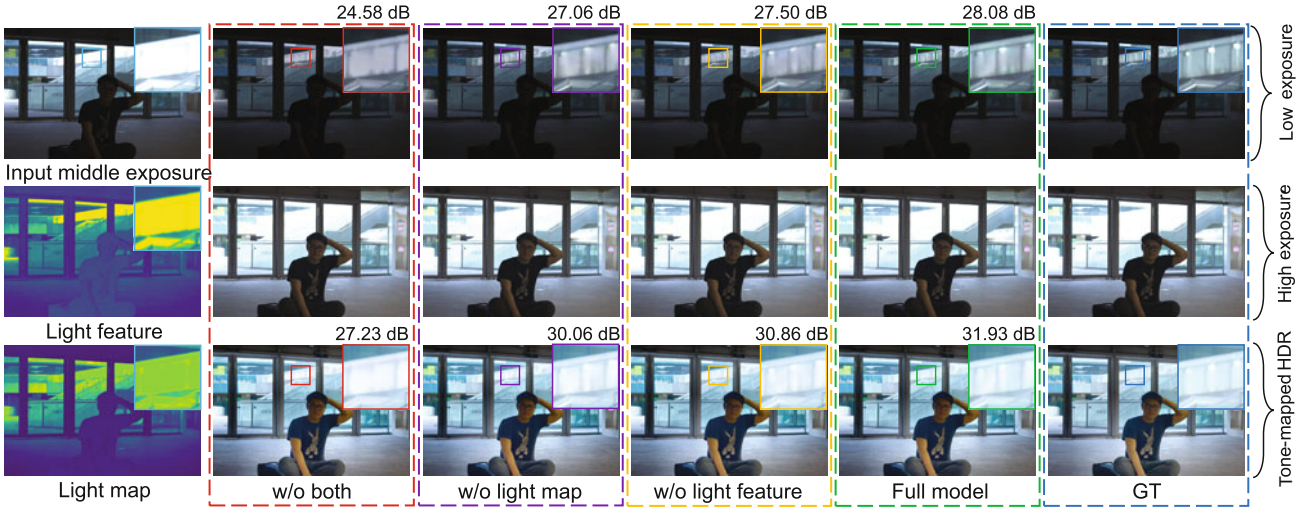

Fig. 10 Visualization and ablation analysis of the luminance estimation module. The estimated light feature and light map are shown together with the generated low- and high-exposure images and the HDR reconstruction results under different ablation settings. Zoomed-in regions and their corresponding PSNR values are provided for clearer comparison. GT: ground truth; w/o: without

Table 5 Impact of different luminance estimation methods on the HDR reconstruction performance over Kalantari's and RealHDRV datasets

Method	Kalantari's dataset				RealHDRV dataset			
	PSNR-μ	PSNR-L	SSIM-μ	SSIM-L	PSNR-μ	PSNR-L	SSIM-μ	SSIM-L
GSB	40.12	37.31	0.9876	<u>0.9793</u>	39.73	32.32	0.9830	0.9877
GSI	40.27	37.56	0.9864	0.9781	40.13	32.15	0.9831	0.9873
HVI-I	40.96	38.08	0.9874	0.9785	40.34	32.54	0.9832	0.9877
RetinexNet	<u>41.28</u>	<u>38.10</u>	<u>0.9878</u>	0.9787	<u>40.73</u>	<u>32.81</u>	**0.9836**	<u>0.9880</u>
USME-HDR**	**41.36**	**38.12**	**0.9879**	**0.9794**	**40.78**	**33.01**	<u>0.9834</u>	**0.9882**

The results indicate that the structured luminance modeling is more effective than simple statistical luminance estimation. Methods marked with ** indicate indirect reconstruction with the 6-channel input X_2. The best and second-best results are highlighted in bold and underlined, respectively

estimation, such as GSB, yields clearly inferior performance on both datasets, indicating that global mean and standard deviation alone are insufficient to describe complex spatial illumination variations. GSI also shows limited improvement, suggesting that low-frequency smoothing alone cannot adequately capture local illumination structure. In contrast, HVI-I, RetinexNet, and our method achieve overall better results, demonstrating that more structured luminance representations are more effective for HDR reconstruction. Among them, the proposed method achieves the best or near-best performance on most metrics, indicating that the learnable luminance estimation module is better suited to the downstream objectives of exposure generation and HDR fusion.

4.6.3 Effect of the exposure time ratio

To investigate the impact of exposure time information on the overall network performance, we conduct an ablation study where the exposure time ratios are excluded from the network input. The quantitative results are reported in Table 4, where the effect of enabling or disabling the exposure time ratio can be observed. Removing exposure time information leads to performance degradation across all evaluation metrics. Fig. 11 presents the visual results with and without incorporating exposure time ratios. It can be observed that integrating exposure time ratios enables more precise exposure control, leading to improved brightness consistency in the generated high- and low-exposure images as well as the final HDR results.

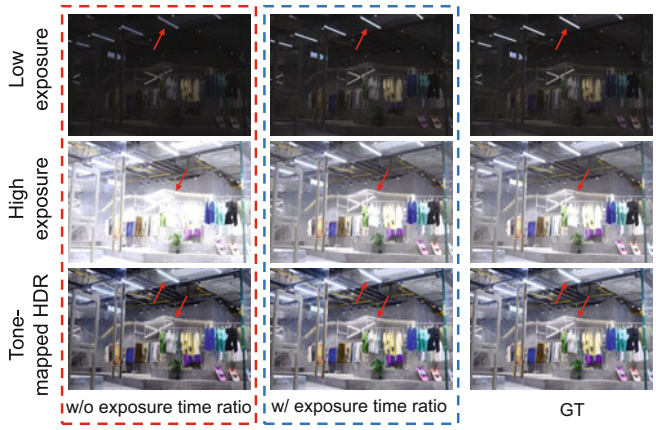

Fig. 11 Visual comparison with (w/) and without (w/o) exposure time ratio. It can be observed that when using the exposure time ratio, the brightness of the generated image is closer to the ground truth (GT)

4.6.4 Effect of the loss function

We assess the contribution of each loss term by progressively incorporating it into the base loss $\mathcal{L}_{\text{sp}} + \mathcal{L}_{\text{se}}$. Table 6 shows that each additional term brings consistent performance gains, and the complete loss formulation achieves the best results, validating the effectiveness of the overall loss design for unsupervised HDR reconstruction.

Table 6 Quantitative results of different loss combinations on the RealHDRV dataset

Loss	PSNR-μ	PSNR-L	SSIM-μ	SSIM-L
$\mathcal{L}_{\text{sp}}+\mathcal{L}_{\text{se}}$	39.07	32.30	0.9667	0.9867
$\mathcal{L}_{\text{sp}}+\mathcal{L}_{\text{se}}+\mathcal{L}_{\text{le}}+\mathcal{L}_{\text{he}}$	40.37	32.43	0.9817	<u>0.9869</u>
$\mathcal{L}_{\text{sp}}+\mathcal{L}_{\text{se}}+\mathcal{L}_{\text{le}}+\mathcal{L}_{\text{he}}+\mathcal{L}_{\text{ea}}$	<u>40.72</u>	<u>32.73</u>	<u>0.9829</u>	0.9868
$\mathcal{L}_{\text{sp}}+\mathcal{L}_{\text{se}}+\mathcal{L}_{\text{le}}+\mathcal{L}_{\text{he}}+\mathcal{L}_{\text{ea}}+\mathcal{L}_{\text{p}}$	**40.78**	**33.01**	**0.9834**	**0.9882**

The best and second-best results are highlighted in bold and underlined, respectively

5 Conclusions

This paper presents USME-HDR, a framework for single-image HDR reconstruction without ground-truth HDR supervision. The proposed method requires only a single LDR image as input at inference time. USME-HDR incorporates an EAN to learn multi-exposure priors, enabling the generation of complementary exposure information from a single input image. Furthermore, inspired by the Retinex theory, we introduce a light map and a light feature to provide luminance-aware guidance for exposure generation, while exposure time ratio guidance further improves luminance fidelity. Finally, the input LDR image is fused with the generated multi-exposure images, and the overall framework is optimized using synthetic pseudo-labels. Experimental results demonstrate that USME-HDR can reconstruct high-quality HDR images from only a single LDR input, without relying on real HDR images or real multi-exposure LDR inputs, highlighting its practical value.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62371175), the Key R&D Program of Zhejiang (No. 2023C01044), and the National Natural Science Foundation of China (No. 62506108).

Author contributions

Han WANG designed the research, performed the investigation, developed the methodology, implemented the software, curated and processed the data, visualized the results, and drafted the paper. Bolun ZHENG acquired the funding, provided resources, supervised the project, and contributed to project administration. Quan CHEN contributed to the investigation and methodology, participated in the project administration, drafted the paper, and revised and edited the paper. Qianyu ZHANG curated and processed the data, conducted the investigation and validation, and contributed to visualization. Tao ZHANG contributed to the conceptualization of the study, provided resources, and supported software development. Jiyong ZHANG and Xiang TIAN revised and edited the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The RealHDRV dataset used in this study is publicly available at https://github.com/yungsyu99/Real-HDRV. Kalantari's dataset is publicly available at https://cseweb.ucsd.edu/~viscomp/projects/SIG17HDR/. The other data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used ChatGPT to improve language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Bai JS, Yin YH, He QY, et al., 2024. RetinexMamba: Retinex-based Mamba for low-light image enhancement. Proc 31st Int Conf on Neural Information Processing, p.427-442. https://doi.org/10.1007/978-981-96-6596-9_30

Cai YH, Bian H, Lin J, et al., 2023. Retinexformer: one-stage Retinex-based Transformer for low-light image enhancement. Proc IEEE/CVF Int Conf on Computer Vision, p.12470-12479. https://doi.org/10.1109/ICCV51070.2023.01149

Chen RF, Zheng BL, Zhang H, et al., 2023. Improving dynamic HDR imaging with fusion transformer. Proc AAAI Conf on Artificial Intelligence, p.340-349. https://doi.org/10.1609/aaai.v37i1.25107

Chen SK, Yen HL, Liu YL, et al., 2023. Learning continuous exposure value representations for single-image HDR reconstruction. Proc IEEE/CVF Int Conf on Computer Vision, p.12944-12954. https://doi.org/10.1109/ICCV51070.2023.01194

Chen XY, Liu YH, Zhang ZW, et al., 2021. HDRUNet: single image HDR reconstruction with denoising and dequantization. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops, p.354-363. https://doi.org/10.1109/CVPRW53098.2021.00045

Chen ZX, Wang YJ, Cai X, et al., 2025. UltraFusion: ultra high dynamic imaging using exposure fusion. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.16111-16121. https://doi.org/10.1109/CVPR52734.2025.01502

Debevec PE, Malik J, 1997. Recovering high dynamic range radiance maps from photographs. Proc 24th Annual Conf on Computer Graphics and Interactive Techniques, p.369-378. https://doi.org/10.1145/258734.258884

Dille S, Careaga C, Aksoy Y, 2024. Intrinsic single-image HDR reconstruction. Proc 18th European Conf on Computer Vision, p.161-177. https://doi.org/10.1007/978-3-031-73247-8_10

Endo Y, Kanamori Y, Mitani J, 2017. Deep reverse tone mapping. *ACM Trans Graph*, 36(6):177. https://doi.org/10.1145/3130800.3130834

Hu T, Yan QS, Qi YK, et al., 2024. Generating content for HDR deghosting from frequency view. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.25732-25741. https://doi.org/10.1109/CVPR52733.2024.02431

Huang X, Zhang Q, Feng Y, et al., 2022. HDR-NeRF: high dynamic range neural radiance fields. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.18377-18387. https://doi.org/10.1109/CVPR52688.2022.01785

Kalantari NK, Ramamoorthi R, 2017. Deep high dynamic range imaging of dynamic scenes. *ACM Trans Graph*, 36(4):144. https://doi.org/10.1145/3072959.3073609

Ke JJ, Wang QF, Wang YL, et al., 2021. MUSIQ: multi-scale image quality Transformer. Proc IEEE/CVF Int Conf on Computer Vision, p.5128-5137. https://doi.org/10.1109/ICCV48922.2021.00510

Khan Z, Khanna M, Raman S, 2019. FHDR: HDR image reconstruction from a single LDR image using feedback network. Proc IEEE Global Conf on Signal and Information Processing, p.1-5. https://doi.org/10.1109/GlobalSIP45357.2019.8969167

Kong LT, Li B, Xiong YK, et al., 2024. SAFNet: selective alignment fusion network for efficient HDR imaging. Proc 18th European Conf on Computer Vision, p.256-273. https://doi.org/10.1007/978-3-031-73347-5_15

Land EH, McCann JJ, 1971. Lightness and retinex theory. *J Opt Soc Am*, 61(1):1-11. https://doi.org/10.1364/JOSA.61.000001

Le PH, Le Q, Nguyen R, et al., 2023. Single-image HDR reconstruction by multi-exposure generation. Proc IEEE/CVF Winter Conf on Applications of Computer Vision, p.4052-4061. https://doi.org/10.1109/WACV56688.2023.00405

Lee S, An GH, Kang SJ, 2018. Deep recursive HDRI: inverse tone mapping using generative adversarial networks. Proc 15th European Conf on Computer Vision, p.613-628. https://doi.org/10.1007/978-3-030-01216-8_37

Li ZW, Zhang F, Cao M, et al., 2024. Real-time exposure correction via collaborative transformations and adaptive sampling. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2984-2994. https://doi.org/10.1109/CVPR52733.2024.00288

Liu SZ, Zhang XD, Sun LC, et al., 2023. Joint HDR denoising and fusion: a real-world mobile HDR image dataset. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.13966-13975. https://doi.org/10.1109/CVPR52729.2023.01342

Liu YL, Lai WS, Chen YS, et al., 2020. Single-image HDR reconstruction by learning to reverse the camera pipeline. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1648-1657. https://doi.org/10.1109/CVPR42600.2020.00172

Liu Z, Wang YL, Zeng B, et al., 2022. Ghost-free high dynamic range imaging with context-aware Transformer. Proc 17th European Conf on Computer Vision, p.344-360. https://doi.org/10.1007/978-3-031-19800-7_20

Mantiuk R, Kim KJ, Rempel AG, et al., 2011. HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans Graph*, 30(4):40. https://doi.org/10.1145/2010324.1964935

Mittal A, Moorthy AK, Bovik AC, 2012. No-reference image quality assessment in the spatial domain. *IEEE Trans Image Process*, 21(12):4695-4708. https://doi.org/10.1109/TIP.2012.2214050

Mittal A, Soundararajan R, Bovik AC, 2013. Making a "completely blind" image quality analyzer. *IEEE Signal Process Lett*, 20(3):209-212. https://doi.org/10.1109/LSP.2012.2227726

Nazarczuk M, Catley-Chandar S, Leonardis A, et al., 2024. Self-supervised HDR imaging from motion and exposure cues. Proc 18th European Conf on Computer Vision, p.363-380. https://doi.org/10.1007/978-3-031-91838-4_22

Prabhakar KR, Senthil G, Agrawal S, et al., 2021. Labeled from unlabeled: exploiting unlabeled data for few-shot deep HDR deghosting. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4873-4883. https://doi.org/10.1109/CVPR46437.2021.00484

Santos MS, Ren TI, Kalantari NK, 2020. Single image HDR reconstruction using a CNN with masked features and perceptual loss. *ACM Trans Graph*, 39(4):1-10. https://doi.org/10.1145/3386569.3392403

Shu Y, Shen L, Hu X, et al., 2024. Towards real-world HDR video reconstruction: a large-scale benchmark dataset and a two-stage alignment network. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2879-2888. https://doi.org/10.1109/CVPR52733.2024.00278

Simonyan K, Zisserman A, 2015. Very deep convolutional networks for large-scale image recognition. Proc 3rd Int Conf on Learning Representations, p.1-14.

Song JW, Park YI, Kong K, et al., 2022. Selective TransHDR: transformer-based selective HDR imaging using ghost region mask. Proc 17th European Conf on Computer Vision, p.288-304. https://doi.org/10.1007/978-3-031-19790-1_18

Tel S, Wu ZW, Zhang YL, et al., 2023. Alignment-free HDR deghosting with semantics consistent transformer. Proc IEEE/CVF Int Conf on Computer Vision, p.12790-12799. https://doi.org/10.1109/ICCV51070.2023.01179

Wang H, Ye M, Zhu X, et al., 2022. KUNet: imaging knowledge-inspired single HDR image reconstruction. Proc 31st Int Joint Conf on Artificial Intelligence, p.1408-1414. https://doi.org/10.24963/ijcai.2022/196

Wang JY, Chan KCK, Loy CC, 2023. Exploring CLIP for assessing the look and feel of images. Proc AAAI Conf on Artificial Intelligence, p.2555-2563. https://doi.org/10.1609/aaai.v37i2.25353

Wei C, Wang WJ, Yang WH, et al., 2018. Deep retinex decomposition for low-light enhancement. Proc British Machine Vision Conf, Article 155.

Wu GY, Fu HM, Liu JY, et al., 2024. Hybrid-supervised dual-search: leveraging automatic learning for loss-free multi-exposure image fusion. Proc AAAI Conf on Artificial Intelligence, p.5985-5993. https://doi.org/10.1609/aaai.v38i6.28413

Xu GW, Wang YJ, Gu JW, et al., 2024. HDRFlow: real-time HDR video reconstruction with large motions. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.24851-24860. https://doi.org/10.1109/CVPR52733.2024.02347

Yan QS, Gong D, Shi QF, et al., 2019. Attention-guided network for ghost-free high dynamic range imaging. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1751-1760. https://doi.org/10.1109/CVPR.2019.00185

Yan QS, Zhang S, Chen WY, et al., 2023a. SMAE: few-shot learning for HDR deghosting with saturation-aware masked autoencoders. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5775-5784. https://doi.org/10.1109/CVPR52729.2023.00559

Yan QS, Chen WY, Zhang S, et al., 2023b. A unified HDR imaging method with pixel and patch level. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.22211-22220. https://doi.org/10.1109/CVPR52729.2023.02127

Yan QS, Yang KZ, Hu T, et al., 2025a. From dynamic to static: stepwisely generate HDR image for ghost removal. *IEEE Trans Circ Syst Video Technol*, 35(2):1409-1421. https://doi.org/10.1109/TCSVT.2024.3467259

Yan QS, Feng YX, Zhang C, et al., 2025b. HVI: a new color space for low-light image enhancement. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5678-5687. https://doi.org/10.1109/CVPR52734.2025.00533

Yang SY, Gu Z, Hao WY, et al., 2025. Few-shot exemplar-driven inpainting with parameter-efficient diffusion fine-tuning. *Front Inform Technol Electron Eng*, 26(8):1428-1440. https://doi.org/10.1631/FITEE.2400395

Yu FH, Gu JJ, Li ZY, et al., 2024. Scaling up to excellence: practicing model scaling for photo-realistic image restoration in the wild. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.25669-25680. https://doi.org/10.1109/CVPR52733.2024.02425

Zhang ZL, Wang HY, Liu S, et al., 2024. Self-supervised high dynamic range imaging with multi-exposure images in dynamic scenes. Proc 12th Int Conf on Learning Representations, p.25867-25884.

Zheng BL, Chen Q, Yuan SX, et al., 2022a. Constrained predictive filters for single image bokeh rendering. *IEEE Trans Comput Imaging*, 8:346-357. https://doi.org/10.1109/TCI.2022.3171417

Zheng BL, Pan XK, Zhang H, et al., 2022b. DomainPlus: cross transform domain learning towards high dynamic range imaging. Proc 30th ACM Int Conf on Multimedia, p.1954-1963. https://doi.org/10.1145/3503161.3547823

Zhu YM, Wang L, Yuan JY, et al., 2025. A ground-based dataset and diffusion model for on-orbit low-light image enhancement. *Front Inform Technol Electron Eng*, 26(7):1083-1098. https://doi.org/10.1631/FITEE.2400261

Zou YH, Yan CG, Fu Y, 2023. RawHDR: high dynamic range image reconstruction from a single raw image. Proc IEEE/CVF Int Conf on Computer Vision, p.12300-12310. https://doi.org/10.1109/ICCV51070.2023.01133