scientific reports

Check for updates

OPEN

A two-stage HDR reconstruction pipeline for extreme dark-light RGGB images

Yiyao Huang[1], Xiaobao Zhu[2], Fenglian Yuan[2], Jing Shi[3]✉, U. Kintak[1]✉, Jingfei Fu[2], Yiran Peng[1] & Chenheng Deng[1]

RGGB sensor arrays are commonly used in digital cameras and mobile photography. However, images of extreme dark-light conditions often suffer from insufficient exposure because the sensor receives insufficient light. The existing methods mainly employ U-Net variants, multi-stage camera parameter simulation, or image parameter processing to address this issue. However, those methods usually apply color adjustments evenly across the entire image, which may cause extensive blue or green noise artifacts, especially in images with dark backgrounds. This study attacks the problem by proposing a novel multi-step process for image enhancement. The pipeline starts with a self-attention U-Net for initial color restoration and applies a Color Correction Matrix (CCM). Thereafter, High Dynamic Range (HDR) image reconstruction techniques are utilized to improve exposure using various Camera Response Functions (CRFs). After removing under- and over-exposed frames, pseudo-HDR images are created through multi-frame fusion. Also, a comparative analysis is conducted based on a standard dataset, and the results show that the proposed approach performs better in creating well-exposed images and improves the Peak-Signal-to-Noise Ratio (PSNR) by 0.16 dB compared to the benchmark methods.

Keywords Extremely dark-light, Image enhancement, Self-attention U-Net, HDR reconstruction pipeline

RGGB array image sensors are widely used in modern smartphones, allowing for high-quality color image capture by detecting red, green, and blue light. This technology is vital for everyday photography and performs exceptionally well in low-light environments, making it ideal for night shots. It has various applications, including digital cameras and high-definition drones. In complex scenes, RGGB images are often combined with depth data to improve object recognition accuracy, essential in fields such as autonomous driving and robotic navigation. Additionally, RGGB sensors are employed in applications like Augmented Reality (AR), Virtual Reality (VR), and Security Monitoring.

High Dynamic Range (HDR) imaging is often limited to specialized hardware, making it costly and impractical for widespread use. With the rise of mobile devices, there is a growing demand for affordable solutions that can capture HDR-quality images. Despite advancements in machine learning for object detection, HDR image detection remains challenging, particularly in low-light conditions where details are often lost. Even human eyes struggle to discern details in dark-light conditions, resulting in underexposed photographs unless camera settings are adjusted to compensate for the lack of light. Retrieving details from dark images is essential for effective processing. Object detectors trained on Standard Dynamic Range (SDR) images struggle under extreme lighting conditions, necessitating enhancements to dimly lit SDR images. Traditional infrared imaging can identify larger objects but fails to recover finer details. Conventional techniques like denoising[1,2] and gamma correction[3] are only partially effective and may misrepresent colors. Recent research has focused on HDR reconstruction from SDR images to improve brightness, which works well for moderately dark images but is less effective in near-darkness. HDR imaging is effective in enhancing slightly underexposed images and revealing details that would otherwise be obscured in shadows. However, its efficacy diminishes significantly when applied to images captured in extremely dark conditions.

To address these limitations, our study aims to enhance the quality of images captured in extremely dark-light conditions, which has been a challenging issue for image processing. The proposed novel approach synergizes the strengths of a self-attention[4] mechanism, a U-net[5] CNN training framework, and a method for linearly

[1]Macau University of Science and Technology, Faculty of Innovation Engineering, Macau 999078, China. [2]Nanchang Hangkong University, School of Information Engineering, Nanchang 330063, China. [3]Department of Mechanical and Materials Engineering, University of Cincinnati, Cincinnati, OH 45221, USA. ✉email: jing.shi@uc.edu; ktu@must.edu.mo

enhanced exposure for HDR image reconstruction. In brief, after subtracting the black level, the input image is first sent to the self-attention U-Net for basic color restoration. After the color correction matrix corrects the possible color cast, enhanced/reduced exposure images are generated through the upper and lower exposure models. After screening the images based on specific criteria, those that meet the requirements are fused to create the final image. This innovative pipeline elevates the brightness of profoundly dark images and augments their clarity, and thus, it outperforms the conventional dark-light processing techniques and CNN models designed for dark-light conditions.

Related works

Much research effort has been dedicated to enhancing images captured in dark light and extremely dark conditions. This includes a variety of approaches, such as traditional image processing algorithms, multi-image HDR reconstruction, single-image HDR reconstruction, and specialized processing pipelines for extremely dark images. However, a critical examination of the datasets employed for training reveals a significant limitation: the majority are categorized under dark-light imaging, predominantly composed of daytime shots. This presents a notable gap in the training material, as such datasets offer limited use for training models to perform under conditions of extreme darkness. The following provides an overview of the existing research in this domain.

Traditional image processing algorithms

In contemporary image enhancement, particularly for dark-light conditions, the predominant denoising methodologies encompass deblurring[1], white balance[6], and gamma correction[7]. After image acquisition, these techniques are integrated into a cohesive processing pipeline tailored to the camera specifications to optimize imaging outcomes. Furthermore, an innovative approach[8] employs pixel categorization, learned linear transformations, and weighted summation to develop an image processing pipeline. Among the noise removal algorithms, those predicate on Nonlinear Total Variation (NLTV)[2] are distinguished for their efficacy in reducing noise while meticulously preserving vital image features. Concurrently, Javier[3] utilizes scale mixtures of Gaussians (SMG) within the wavelet domain, a technique that synergizes the strengths of wavelet transforms with statistical models to denoise images proficiently. Moreover, a substantial corpus of research[9,10] has been dedicated to single-image denoising techniques, which predominantly leverage the inherent prior information of images, such as smoothness, sparsity, dark rank, or self-similarity, as the foundation for noise reduction strategies.

Additionally, the utilization of multiple images for denoising, particularly in continuous shooting scenarios, has been explored. Techniques such as Fast Burst Image[11], Deep Burst Denoising[12], and Kernel Prediction Networks[13] focus on rapidly capturing a sequence of images with a camera. Traditional image denoising methods are limited by their tendency to be over-smooth, resulting in the loss of critical details like edges and textures and unnaturally flat images. They often fail to distinguish noise from high-frequency components, inadvertently removing vital features. The high computational cost and potential for introducing new artifacts further compromise image quality.

Multi-image HDR reconstruction

The synthesis of HDR images from multiple exposures is a pivotal technique in overcoming the limitations of conventional imaging sensors, which often struggle to capture the full spectrum of light in scenes with varying brightness levels. This process combines images taken at different exposure levels to create a single image that accurately represents a broader range of luminance levels. The photography pipeline proposed by Hasinoff et al.[14] is claimed to be effective in enhancing image quality by sequentially capturing, aligning, and merging frames. This process results in intermediate images with increased bit depth, dynamic range, and reduced noise compared to the original frames. Yan et al. propose a multi-scale architecture and a residual network[15] to reconstruct HDR images, employing a coarse-to-fine strategy. Their method focuses on learning the relative changes between input frames and ground truth, generating artifact-free images, and restoring missing information.

Diverging from traditional dark-based approaches, a novel non-streaming depth framework[16] is proposed for HDR imaging in dynamic scenes, and is particularly effective in handling large-scale foreground motion. The method conceptualizes HDR imaging as an image translation issue, circumventing the need for extreme dark light scenes. It is shown that the translation network can produce authentic HDR details even in challenging conditions such as complete occlusion, saturation, and underexposure. Similarly, a learning-based solution for HDR imaging in dynamic scenes[17] is developed to utilize CNNs to model the HDR merging process. The study compares three distinct system architectures—Direct, Weight Estimator, and Weight and Image Estimator—to generate high-quality HDR images from sets of three dark SDR images. In brief, multi-image HDR reconstruction techniques, which combine multiple exposures, have the potential to improve Peak-Signal-to-Noise Ratio (PSNR) metrics but struggle in extremely dark conditions. Those methods require images at varying exposures, posing challenges for real-time applications. Precise alignment in dynamic scenes can introduce artifacts, and high computational demands limit real-time use.

Single image HDR reconstruction

Single-image HDR reconstruction techniques have emerged as a powerful solution to overcome the limitations of capturing the full dynamic range of real-world scenes with standard cameras. These methods focus on reversing the camera imaging pipeline, transforming an SDR image into an HDR image by inferring the missing details in under or over-exposed regions. A novel pipeline[18] deconstructs the HDR to LDR conversion process into three distinct stages: dynamic range clipping, nonlinear mapping through the camera response function (CRF), and quantization. Deep neural network model[19] that simplifies finding mappings between LDR and HDR images of different bit depths by inferring LDR images of the same scene under various conditions. DRTMO[20]

employs supervised learning to synthesize SDR images captured under diverse exposures, which are then used to reconstruct and merge HDR images. An enhancement algorithm[21] utilizing a cross-bilateral filter processes the spatial domain and luminance range, employing a Gaussian equation as a descent function to effectively manage exposure problems. Inverse tone mapping algorithm inspired by the retinal response of the human visual system[22] grounded in physiological research, this approach circumvents the artifact issues prevalent in many existing methods. Le et al.[23] generate multiple exposure images by inversely learning the physical image formation process and then fuse these images to generate the HDR image.

While single-image HDR reconstruction can rapidly and effectively enhance SDR images under dark-light conditions, its performance, as indicated by improvements in PSNR values, does not measure up to that of multi-image HDR techniques. In scenarios involving particularly dark images, single-image HDR reconstruction's efficacy is constrained, frequently yielding images that remain dark after processing.

Learning-based pipelines for low-light image processing

Deep learning methods are usually used to achieve effective results in reconstructing extremely dark images. For instance, the two methods of See in the Dark (SID)[24] and See Motion in the Dark (SMID)[25] enhance the quality of images captured in very dark-light conditions, addressing noise and color cast issues. Both methods construct a dark-light image dataset and train an end-to-end fully convolutional neural network using pairs of short- and long-exposure images. Cai et al.[26] propose a method leveraging an improved U-Net architecture incorporating recursive residual convolution units (RRCUs) and dilated convolutions. This approach modifies the standard convolutional blocks and pooling operations to enhance feature extraction and image reconstruction.

It can also be reconstructed by camera sensor modeling. DeepISP[27] represents a shift towards direct processing of raw sensor data through deep learning, aiming to generate visually superior images in a single step. This reduces manual intervention and improves performance by learning the entire image signal processing pipeline. In[28], the residual learning-based method for denoising images in extremely dark-light environments simplifies noise removal by focusing on learning the differences between noisy and clean images. Despite challenging lighting, it aims to produce more precise and visually appealing results. Zamir et al.[29] develop a digital camera pipeline optimized for extreme dark light imaging, in which the learning-based model adopts traditional camera pipeline stages for minimal lighting scenarios, such as demosaicing, denoising, and color correction. The pipeline enhances image quality in near-darkness by utilizing machine learning, addressing a critical need in photography and surveillance for high-quality images in poorly lit environments. Cao et al.[30] propose a physically based ISO-dependent sensor noise model to more accurately simulate camera noise in extreme low-light conditions.

There are also many other ways to model the color. Dong et al.[31] generate monochrome raw images through a deep neural network and then achieve low-light image enhancement by fusing color raw data and a channel attention mechanism. This method utilized the fusion of virtual monochrome and color raw images to improve image quality significantly. Jin et al.[32] introduce the Decouple and Feedback Network (DNF), a novel framework designed to address challenges in low-light image enhancement. The DNF framework strategically decouples the enhancement process into two domain-specific tasks: denoising in the RAW domain and color restoration in the sRGB domain. These are achieved by applying advanced denoising algorithms that preserve fine details and prevent the loss of low-light image characteristics.

Image restoration methods

Bhandari et al.[33] employ a multi-exposure histogram equalization model to improve image quality by optimizing contrast and brightness. This technology processes multiple exposure images to create a composite image that maximizes brightness and contrast without losing detail. Subramani et al.[34] proposed a novel Bezier curve modification method, which uses the Bezier curve to adjust the contrast of images, especially for images with degraded contrast. By optimizing the control points of the Bezier curve, the visual effect of the image can be effectively improved. Subramani et al.[35] select the best enhancement parameters through the cuckoo search algorithm and use its good global convergence ability and local search ability to improve the contrast and details of the image. Veluchamy et al.[36] combine fuzzy dissimilarity metric, contextual intensity transform, and gamma correction to improve the quality of color images. First, the fuzzy dissimilarity metric is used to evaluate the differences between different regions in the image. Then, the contextual intensity transform dynamically adjusts the intensity value according to local features. Gamma correction technology is then combined further to optimize the brightness and contrast of the image.

Datasets used in literature

The RENOIR[37] and DND[38] datasets are for the same scene that takes low ISO images as ground truth and high ISO images as noise images. They adjust camera parameters such as exposure time to make the two images have the same brightness. Most of the images of the datasets used in Nam[39], PolyU[40], and SIDD[15] are synthesized. There is a distinct difference between them and the authentic noisy images, which limits the denoising effect of the network trained on this data set on authentic noisy images. This study uses the extreme dark images dataset[24]. The dataset contains the original collection of extremely dark light images of RGGB using cell phone lenses, each with a corresponding bright and clear reference image that can be used as a training set. This dataset has been published since 2018, and many studies using this dataset have been published in recent years in premium journals[27,41] and proceedings of computer vision conferences such as CVPR[24,30,32] and ICCV[25].

Methodology

Problem statement

In image processing, HDR reconstruction techniques have proven highly effective in improving low-light images, as evidenced by the example shown in Fig. 1. These techniques widen the dynamic range of images, revealing details that may be hidden in shadows or overexposed areas. The fundamental concept behind HDR reconstruction involves combining multiple images taken at different exposure levels, collectively capturing a wider range of luminance than a single exposure can achieve. This approach has been crucial in overcoming the limitations of traditional imaging sensors, allowing for the retrieval of details from the darkest and brightest parts of a scene.

However, applying HDR reconstruction techniques faces a significant challenge when dealing with extremely dark images, as shown in Fig. 1. In these instances, the limited light makes it hard to see details, hindering the usual HDR reconstruction process. The main problem is the insufficient luminance data in the images, which is essential for combining multiple exposures into a cohesive HDR image. As a result, the typical HDR method, which relies on merging differently exposed images, is ineffective in such extreme dark-light conditions.

A new approach is necessary to overcome this limitation, which goes beyond the traditional HDR reconstruction paradigm. Our proposal involves harnessing machine learning techniques to aid in reconstructing extremely dark images. Training a neural network on a diverse dataset of images captured under various lighting conditions enables the model to infer missing details from severely underexposed images. This learning-based reconstruction method has the potential to extract discernible information from images that would otherwise be deemed irrecoverable, thereby broadening the applicability of HDR reconstruction techniques to a wider range of challenging lighting scenarios.

Our research aims to expand on the current foundation by introducing innovative methodologies to further enhance the effectiveness of HDR reconstruction techniques, especially in scenarios involving extremely dark images. By incorporating advanced machine learning algorithms and utilizing cutting-edge neural network architectures, our work addresses the limitations of traditional HDR reconstruction methods. The goal is to significantly improve the ability to recover detailed information from images characterized by severe underexposure, thereby extending the applicability and performance of HDR reconstruction in extremely dark-light conditions.

The proposed two-stage pipeline is crucial for restoring images captured in extremely dark lighting conditions. Initially, the image exposure is adjusted to achieve an optimal brightness level. Following this initial correction, the focus shifts to enhancing the image sharpness as much as possible. To achieve this, the proposed pipeline utilizes a neural network specifically trained to enhance the luminance of underexposed images. Subsequently,

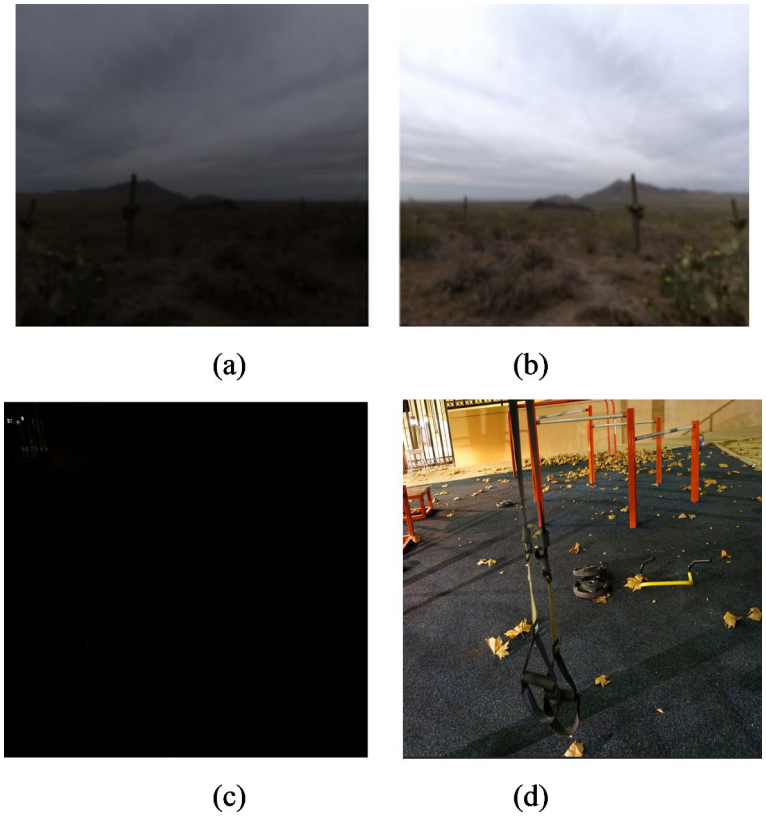

Fig. 1. Examples of (**a**) low light image, (**b**) ground truth of the low light image, (**c**) extreme dark image, and (**d**) ground truth of the extreme dark image. The low-light image is blurred and visible, while the extreme-dark light raw image is indistinguishable.

we integrate advanced models designed to refine the clarity of the image, resulting in a sharper and more detailed visual output.

Proposed pipeline structure

The image processing workflow, depicted in Fig. 2, begins with the input of an image into the self-attention mechanism and U-Net CNN architecture. This initial stage produces an SDR image with black levels eliminated. The image then undergoes white balancing through the application of color correction matrix (CCM). Thereafter, the camera response curve is used to adjust the image exposure, resulting in multiple images with linearly enhanced exposure levels. These images are then merged through a selection process governed by a specific algorithm. This computational process ultimately yields an HDR image.

To ensure that the enhancement effect will not be compromised without a feedback mechanism, we have implemented the following considerations. First, the self-attention U-Net in the framework focuses on rectifying the incomplete aspects of the black level removal method during the preprocessing phase after reading the RGGB images. This stage aims to restore the main content of the image. In our preliminary test, after 4,000 iterations, the self-attention U-Net network achieved a loss of less than 0.02 (please refer to "Experiments" section for the loss diagram and related training process details). Therefore, the basic quality of images is guaranteed. Second, the first evaluation and adjustment involve the color correction matrix (described in "Color correction matrix" section). Our preliminary tests indicated that images generated by the U-Net network sometimes exhibited a color cast. The color correction matrix is adopted to restore and correct any color offsets to address this issue. If the U-Net network generates a properly balanced image, the colors will remain accurate after this operation. Third, the second evaluation and adjustment involve the pixel exposure enhancement screening algorithm (as outlined in "Pixel exposure difference algorithm" section). Adjustments made for exposure can sometimes lead to issues with overexposure or underexposure. To ensure image quality, we utilize the differences between two adjacent pixels to determine which images are suitable for fusion, ultimately producing the final output. Therefore, even without a feedback mechanism, our pipeline includes functional corrections that help minimize errors that could negatively affect the final results.

Self-attention U-Net network

Self-attention U-Net architecture

Figure 3 illustrates the schematic of the proposed self-attention U-Net model. Within this model, the input image undergoes a systematic filtering and down-sampling process, reducing by a factor of 2 at each scale within the network's encoding segment. The Attention Gates (AGs) are crucial in filtering features transmitted through skip connections. Within the AGs, the selective processing of features is enabled by utilizing contextual information extracted at coarser scales, known as gating. We introduce a novel grid-based gating approach that allows attention coefficients to be specifically targeted towards local regions. This methodological innovation offers significant performance improvement compared to traditional gating mechanisms that rely on global feature vectors. By extending the standard U-Net model, we seek to enhance the model sensitivity to foreground pixels, thus eliminating the need for complex heuristic approaches.

It is expected that the self-attention mechanism can extract more effective features from the main body of an image in this scenario to help restore the main body of the image. However, it could affect the quality of generated images in some indoor scenes. In this regard, the subsequent step (multi-frame HDR fusion) can mitigate this problem by effectively improving the brightness of the images. It can generate a sufficient number of enhanced exposure images for fusion, and the existence of the screening mechanism avoids overexposure.

In convolutional neural networks, local information processing is meticulously orchestrated across successive layers to distill higher-dimensional representations of an image incrementally, symbolized as x^l. This intricate procdure effectively discriminates pixels within the expansive high-dimensional space, categorizing them according to their semantic attributes. The predictive prowess of the model is intricately linked to the synthesis of information amassed from an extensive receptive field facilitated by this methodical progression. As a direct consequence, the feature map x^l emerges at the output of layer , realized through the methodical application of linear transformations followed by nonlinear activation functions. The Rectified Linear Unit (ReLU), denoted conventional as $\sigma_1\left(x^l_{i,c}\right) = \max(0, x^l_{i,c})$ —with and csignifying the spatial and channel dimensions, respectively—is the activation function of choice. The activation of features is succinctly represented as $\frac{1}{x_c} = \sigma_1\left(\sum_{c' \in F_l} x^{l-1}_{c'} * k_{c',c}\right)$, where the asterisk (*) signifies the convolution operation, and the spatial subscript is deliberately omitted to enhance notational clarity. The defining attribute

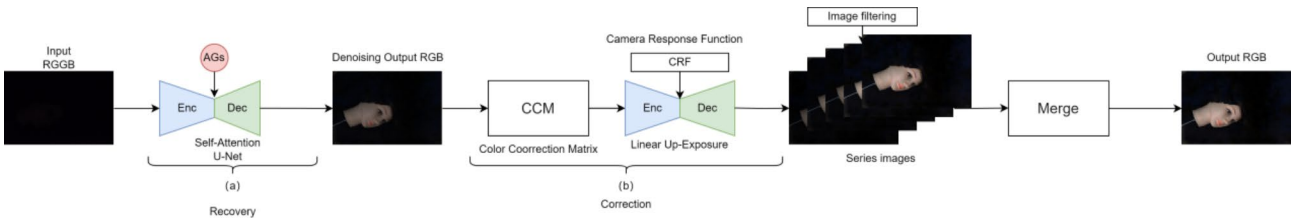

Fig. 2. Schematic of the proposed pipeline, consisting of sequential steps of self-attention mechanism, U-Net CNN network, color correction, image filtering, and fusion.

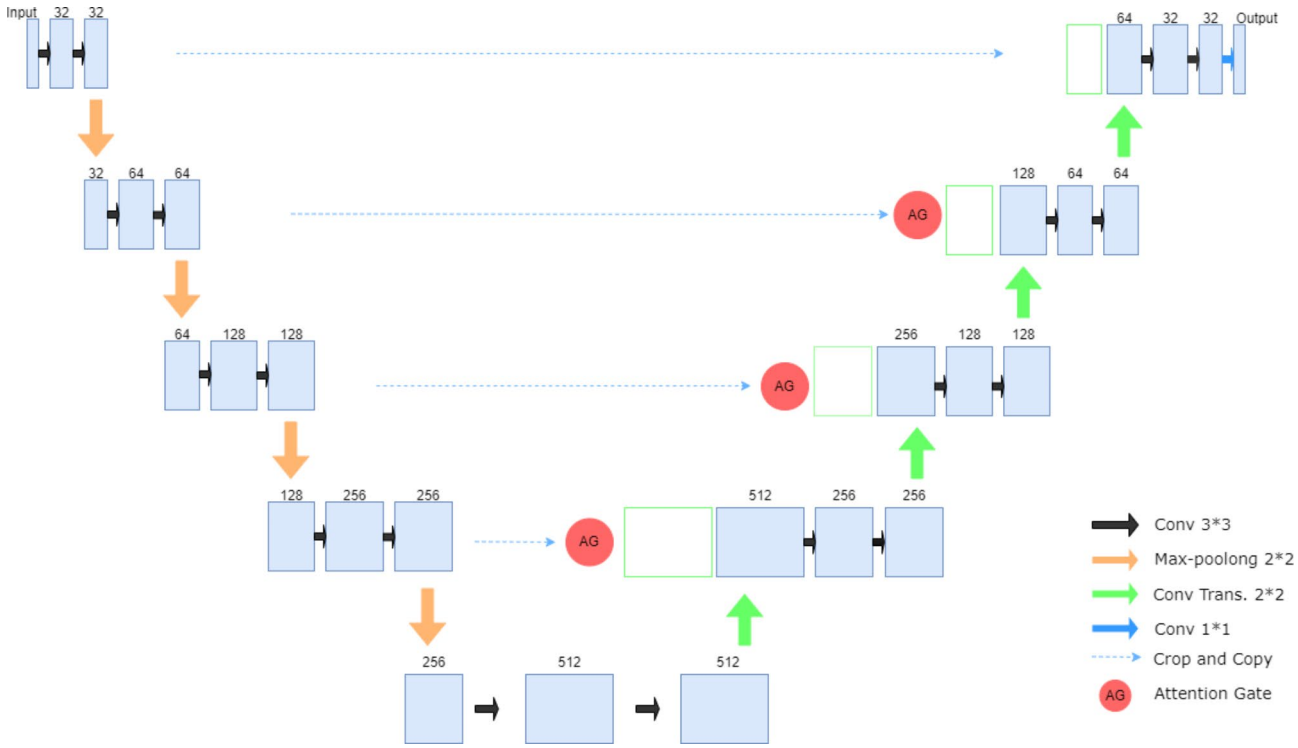

Fig. 3. Schematic of self-attention U-Net architecture, characterized by Attention Gate modules added to all jump links to reduce redundant Skip Connection instead of directly connecting features from the same down-sampled layer to the up-sampled layer.

of the function $f(x^l; \Phi^l = x^{l+1})$, applied at the convolutional layer , is encapsulated by the trainable kernel parameters Φ^l. These parameters are meticulously refined by minimizing the training objective via Stochastic Gradient Descent (SGD).

Within the scope of this paper, we introduce a self-attention model that innovatively extends the canonical 4-layer U-Net architecture, imbuing it with enhanced focus and discriminative capabilities. This model is a testament to the potential of integrating self-attention mechanisms within established architectures, promising advancements in image enhancement, particularly in scenarios characterized by extremely dark-light conditions.

Attention gate

In the proposed framework, the attention coefficient $\alpha_i \in [0, 1]$ plays a pivotal role in discerning salient regions within the image, effectively pruning the feature responses to preserve only those activations that are pertinent to the task at hand. The AG output is derived from an element-wise multiplication of the input feature map with the attention coefficient, denoted as $\hat{x}^l_{i,c} = x^l_{i,c} \cdot \alpha^l_i$. In our standard configuration, a singular scalar attention value is computed for each pixel vector $\mathbf{x_i^l} \in R^{F_l}$, where F_l represents the number of feature maps at layer . Each AG is adept at learning to concentrate on a specific subset of target structures within the image. As depicted in Fig. 4, a gating vector $\mathbf{g_i} \in R^{F_g}$ is employed for each pixel $i$to ascertain the focal area. This gating vector is imbued with contextual information that is instrumental in pruning the responses of lower-level features. To calculate the gating coefficients, we utilize an additive attention mechanism[42], which, despite its higher computational demand, has been demonstrated through experiments to yield superior accuracy compared to its multiplicative counterpart[43]. We use a medical imaging attention gate[44], the summation is computed as,

$$q^l_{att} = \Psi^T \left(\sigma_1 \left(W_x^T x_i^l + W_g^T g_i + b_g\right)\right) + b_\Psi \tag{1}$$

$$\alpha_i^l = \sigma_2 \left(q^l_{att} \left(x_i^l, g_i; \Theta_{att}\right)\right) \tag{2}$$

where $\sigma_2 \left(x_{i,c}\right) = \frac{1}{1+exp(-x_{i,c})}$ corresponds to the sigmoid activation function, effectuating a linear transformation through a channel-wise $1 \times 1 \times 1$ convolution of the input tensor. This sophisticated attention mechanism ensures that our model enhances the overall image quality and preserves the integrity of the scene's textual information, even under extremely dark-light conditions, as corroborated by the state-of-the-art performance on benchmark datasets. The operational mechanism of AGs is further explained in Fig. 4.

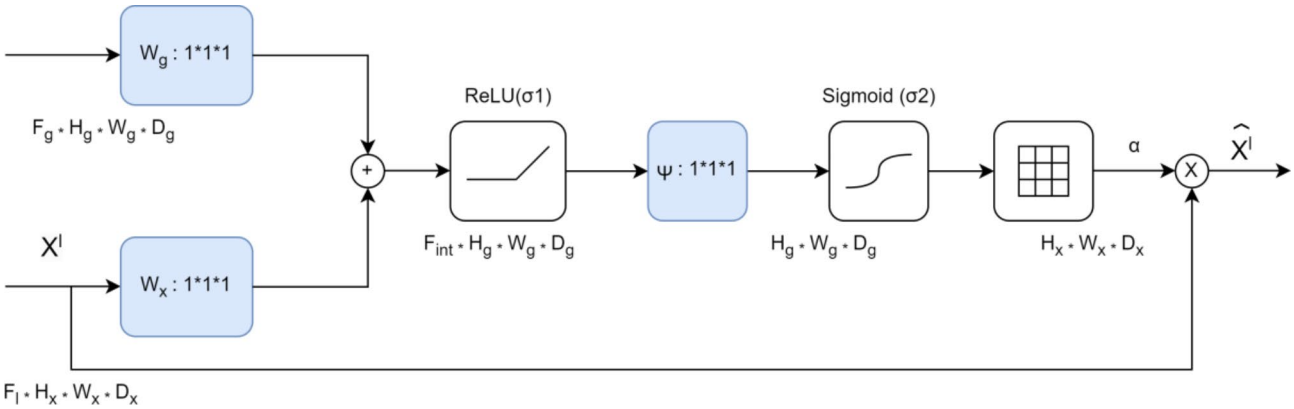

Fig. 4. Schematic of self-attention gate.

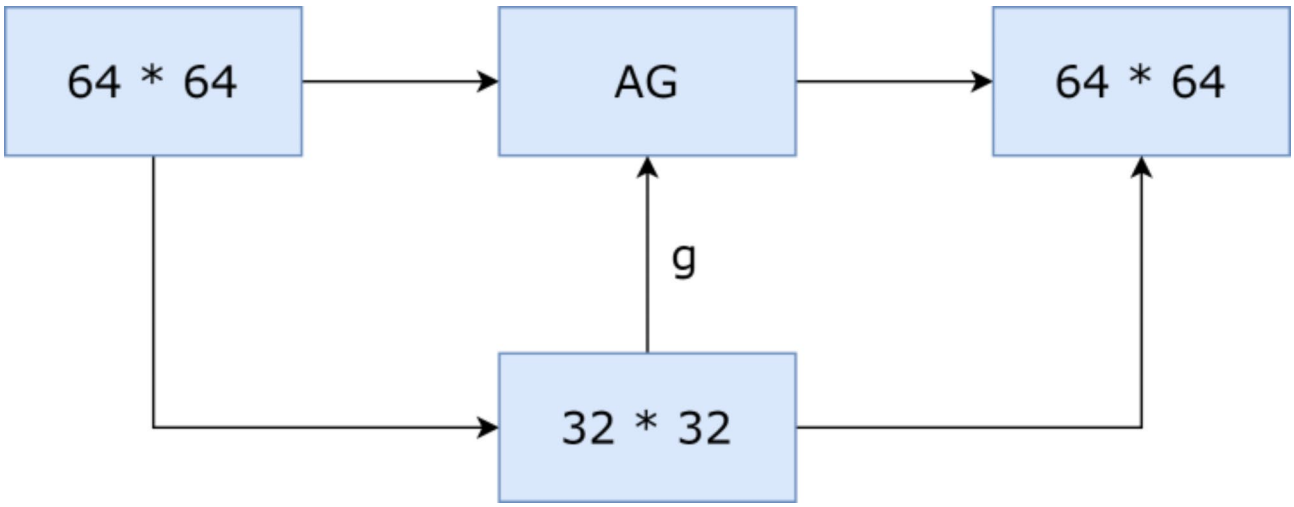

Fig. 5. An example on how AGs are implemented at every skip connection.

Attention gate implementation at the skip connection

An example of how AGs are implemented at every skip connection is shown in Fig. 5. The attention gate operates by integrating two input vectors, $\mathbf{x}$ and $\mathbf{g}$, where $\mathbf{g}$ is sourced from the subsequent lower layer of the network and possesses smaller dimensions alongside superior feature representation due to its derivation from a deeper network level. For instance, in the illustrated example, vector $\mathbf{x}$ would exhibit 4 * 64 * 64(filters * height * width), while vector $\mathbf{g}$ would present 2 * 32 * 32. Vector $\mathbf{x}$ undergoes a stride convolution, adjusting its dimensions to 64 * 32 * 32, and concurrently, vector $\mathbf{g}$ is processed through a 1 * 1 convolution, resulting in matching dimensions of 64 * 32 * 32. Following this, the two vectors are subjected to an element-wise summation, which amplifies aligned weights and diminishes the relative magnitude of unaligned weights. The emergent vector is then channeled through a ReLU activation layer and a 1 * 1 convolution, which reduces its dimensions to 1 * 32 * 32. Subsequently, this vector is passed through a sigmoid layer that normalizes its values to the range, thereby generating the attention coefficients (weights) that signify feature relevance, with coefficients approaching 1 denoting higher relevance. These attention coefficients are then up-sampled back to the original dimensions of vector $\mathbf{x}$ (i.e., 64 * 64) via bilinear interpolation. The final step involves an element-wise multiplication of the attention coefficients with the original $\mathbf{x}$ vector, effectively scaling the features by their relevance. The refined $\mathbf{x}$ vector, modulated by the attention mechanism, is propagated through the skip connection as usual in the network architecture.

We utilize the proposed AGs within the standard U-Net architecture to emphasize salient features transmitted through skip connections. Information extracted from coarser scales is employed for gating, eliminating ambiguities in irrelevant and noisy responses within the skip connections. This operation is conducted before sequential processing, ensuring only pertinent activations are merged. Moreover, AGs filter neuronal activations during both forward and backward propagation. Gradients originating from background regions are assigned reduced weights during the backward pass, allowing model parameters in the shallower layers to be updated primarily based on spatial areas relevant to the given task. The update rule[42] for the convolutional parameters at layer $l-1$ can be articulated as follows:

$$\frac{\partial(\widehat{x_i^l})}{\partial(\varphi^{l-1})} = \frac{\partial(\alpha_i^l f(x_i^{l-1};\varphi^{l-1}))}{\partial(\varphi^{l-1})} = \alpha_i^l \frac{\partial(f(x_i^{l-1};\varphi^{l-1}))}{\partial(\varphi^{l-1})} + \frac{\partial(\alpha_i^l)}{\partial(\varphi^{l-1})} x_i^l \tag{3}$$

In the context of attention mechanisms within U-Net for enhancing extremely low-light images, the first gradient term on the right-hand side is scaled by α_i^l. In the case of multi-dimensional AGs, α_i^l corresponds to the vector at each grid scale. Within each sub-AG, complementary information is extracted and fused to define the output of the skip connections. To reduce the number of trainable parameters and the computational complexity of the AGs, linear transformations are performed without any spatial support ($1\times1\times1$ convolution), and the input feature map is downsampled to the resolution of the gating signal, akin to non-local blocks. The corresponding linear transformations decouple the feature mappings and project them into a lower-dimensional space for gating operations. Our network omits the first connection from the gating function, as it does not represent the input data in the high-dimensional space. We employ deep supervision to enforce the semantic distinctiveness of the intermediate feature maps at each image scale. This ensures that attention units at different scales can influence responses to the extensive foreground content of the image. Consequently, we prevent dense predictions from being reconstructed from a small subset of skip connections.

Overall process

Figure 6 delineates the workflow applied to the images that have undergone training. The dataset comprises original RGGB raw images with 4240×2832 pixels dimensions. Initially, the RGGB images are divided into four distinct channels. Subsequently, the black level is subtracted, and the data are scaled by a factor of 100. A segment of the image, measuring 512×512 pixels, is then extracted and input into our enhanced training network. The image segment is subjected to random flipping and transposition to augment the network's robustness and mitigate the risk of overfitting.

Color correction matrix

Once the network generates the results, the color output of the image sensor is adjusted using a Color Correction Matrix (CCM) to ensure that the captured colors accurately reflect the colors of the actual scene. The CCM is represented as a 3×3 matrix, where each element is employed to modify the original color values (red, green, and blue) to obtain the corrected color values.

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} \boldsymbol{CC}_{11} & \boldsymbol{CC}_{12} & \boldsymbol{CC}_{13} \\ \boldsymbol{CC}_{21} & \boldsymbol{CC}_{22} & \boldsymbol{CC}_{23} \\ \boldsymbol{CC}_{31} & \boldsymbol{CC}_{32} & \boldsymbol{CC}_{33} \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{4}$$

where R, G, B are the original color values, R', G', B' are the corrected color values, and $\boldsymbol{CC}_{ij}$ are the coefficients of the matrix that adjust the contribution of each color channel. This adjustment addresses color distortions caused by the learned loss, including white balance correction, removal of color shifts, and enhancement of image brightness and contrast.

Linearly enhanced exposure model

Linearization aims to estimate the Camera Response Function (CRF) and transform nonlinear SDR images into linear irradiance maps. There are numerous established methods in existing literature for generating sets of linearly exposed images. We employ the over/under exposure networks from DRTMO[20] to develop two sets of images with varying exposures. The CRFs are derived using the DoRF[45] response functions, from which five representative curves are obtained via the k-means method and subsequently normalized. Different nonlinear camera response curves are utilized to alter the exposure of images, resulting in multiple images with varied exposure levels. The dataset is thus composed of SDR images as inputs and sets of images with different exposure levels as outputs. The calculation method is as follows:

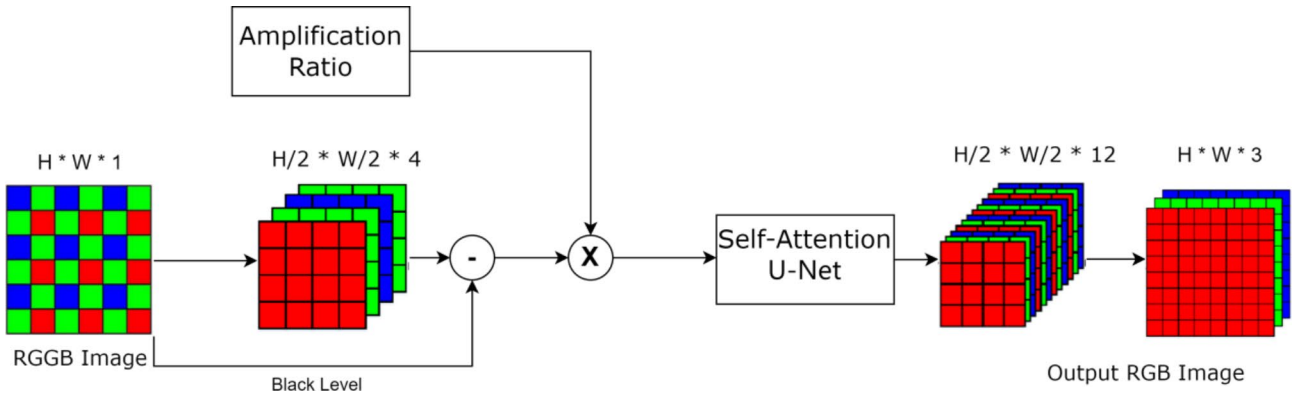

Fig. 6. Proposed image processing pipeline. The input image is divided into four different channels, and after subtracting the black level, a portion of 512 * 512 image is extracted and fed into our self-attention U-Net.

$$Z_{i,j} = f(E_i \Delta t_j) \tag{5}$$

$$\Delta t_j = \frac{1}{\tau^{T/2}}, \ldots, \frac{1}{\tau^2}, \frac{1}{\tau}, 1, \tau, \tau^2, \ldots, \tau^{T/2} \tag{6}$$

where $Z_{i,j}$ denotes the pixel value at point i in the SDR image under exposure index j, represents the camera response function, E_i signifies the pixel value at point i in the HDR image, and Δt_j stands for the exposure time. This formulation encapsulates the relationship between the pixel values of SDR and HDR images through the lens of the camera's response to varying exposure levels. Specifically, the camera response function serves as a bridge, translating the SDR pixel values $Z_{i,j}$ at a given exposure index j and exposure time $\Delta \mathrm{t_j}$ into the HDR domain, culminating in the pixel value $\mathrm{E_i}$at the corresponding point i. This interplay is pivotal in reconstructing the high dynamic range scene from its low dynamic range counterparts, leveraging the exposure time as a critical parameter.

Encoding a single SDR image into the depths of a neural network is achieved through a 2D convolutional neural network. Subsequently, a 3D convolutional neural network decodes the deep semantic features of the image into SDR images of different exposure levels. This network architecture is designed to adjust the exposure of images, either increasing or decreasing them, to yield multiple images that will later be synthesized into an HDR image. This approach facilitates the generation of images with a wide range of exposures and enhances the subsequent HDR composition by providing richer information for the merging algorithm. Since the model requires extensive training time and resources, a pre-trained model is adopted and verified after testing because its performance generally meets the experimental requirements. Consequently, the SDR images obtained from the previous step are input into the model, which yields two sets of images with varying exposure levels through the over/under exposure model application. This approach streamlines the process by leveraging existing resources and efficiently generates the desired output, facilitating further experimentation with images of different exposure degrees.

Note that the implementation of 3D CNNs is expected to impact efficiency to some extent. Our primary focus is on enhancing image quality - resource consumption will be considered unless subsequent steps can effectively offset the resource losses incurred from earlier operations. Second, significant color distortion or noise in the initial step could affect the following stages. However, the self-attention U-Net can successfully restore the main content of the image, resulting in minimal loss after 4000 iterations. In our experiment, no significant distortion is discovered, and the issue of color cast can be effectively addressed using a color correction matrix.

Pixel exposure difference algorithm

Upon feeding SDR images into the Encoder-Decoder neural network, the overexposure network outputs N overexposed images, while the underexposure network produces N underexposed images. This process results in a collection of $2\mathrm{N} + 1$ SDR images with varying exposure levels, from which we select K images to synthesize into a High Dynamic Range (HDR) image. However, if the input SDR image is overexposed or underexposed, the generated SDR images may contain erroneous pixels in the bright or dark regions, respectively. To circumvent these imperfections, we employ the following heuristic approach:

$$|v_{j+1} - v_j| < \eta \tag{7}$$

Starting with the input image, we progressively select the $(j+1)^{th}$ image that is brighter or darker until the pixel value v_{j+1} for each channel is either greater or less than v_j, or the relative difference between them reaches a predefined threshold—following testing. The threshold value is set to 64 to ensure effective screening for various exposures, avoid prolonged compositing times due to excess compostable images, or prevent loss of detail from insufficient images. This heuristic retrieval ensures that the selected images for HDR synthesis are free from flawed pixels that could compromise the quality of the final HDR image, thereby maintaining the integrity of the image's luminance details.

Multi-image fusion algorithm

Fusion process

In the final step of our process, we aim to merge the selected images into the ultimate HDR image. Traditionally, direct synthesis requires the exposure time for each bracketed image. However, we adopt the Exposure Fusion[46] to minimize computational demands and facilitate comparison. This approach does not necessitate exposure time as an input. Instead, it generates a tone-mapped SDR image by directly fusing multiple exposures into a high-quality image with reduced dynamic range. The core methodology involves calculating a perceptual quality metric for each pixel within the multi-exposure sequence. This metric encodes the desired qualities, such as saturation and contrast. Based on our quality assessment criteria, we select the "best" pixels from the sequence and amalgamate them into the final composite. This entire procedure is depicted in Fig. 7, showcasing a streamlined approach to achieving high-quality HDR imagery without the complexities associated with traditional HDR photography methods.

Fusion method

In generating a series of images for HDR composition, some images will inevitably need more underexposure or overexposure, resulting in many images within the stack containing flat, desaturated areas. Such regions, devoid of weight, should be excluded, while areas rich in vibrant colors and details must be preserved. To achieve this, we employ three metrics. Firstly, a Laplacian filter is applied to each image's grayscale version, and the filter response's absolute value is obtained, producing a contrast metric C that assigns higher weights to significant

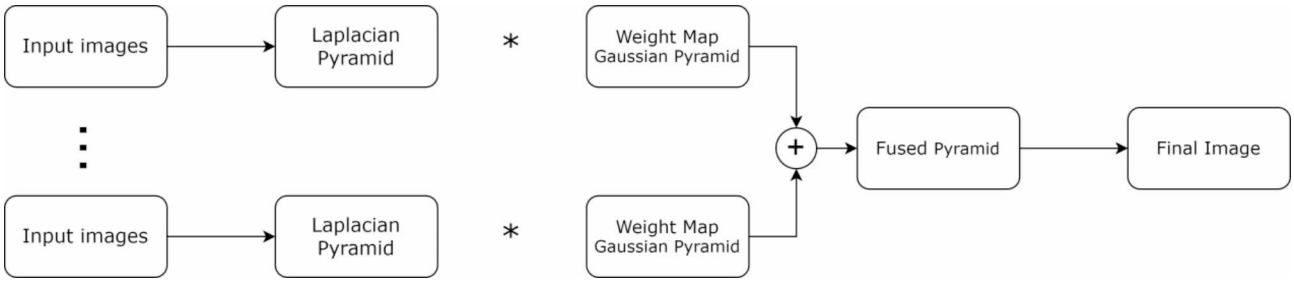

Fig. 7. Fusion exposure algorithm that skips calculating HDR and directly fuses multiple exposures into a high-quality, low-dynamic range image.

elements like edges. Secondly, as the exposure time of a photograph increases, the resulting colors become desaturated and eventually clipped. Saturated colors are desirable and render the image more vivid. A saturation metric S is calculated at each pixel as the standard deviation within the R, G, and B channels. Lastly, examining the raw intensities within the channels reveals the extent of a pixel's exposure. We aim to retain intensities that are neither close to 0 (underexposed) nor 1 (overexposed). Utilizing a Gaussian curve $exp\left(-\frac{(i-0.5)^2}{2\sigma}\right)$, where $\sigma = 0.2$, each intensity is weighted according to its proximity to 0.5. The Gaussian curve is applied to each channel individually, and the results are multiplied to derive the measure of well-exposedness E.

For each pixel, we amalgamate information derived from various metrics into a scalar weight map through multiplication, akin to the weighted terms in a linear combination. To modulate the influence of each metric, we employ a power function. This methodology facilitates a sophisticated integration of diverse image attributes, ensuring a balanced and precise image enhancement based on a comprehensive assessment of its characteristics.

$$W_{ij,k} = (C_{ij,k})^{W_C} * (S_{ij,k})^{W_S} * (E_{ij,k})^{W_E} \tag{8}$$

where the subscript ij, k refers to the pixel located at position (i, j) in the k^{th} image.

We compute a weighted average along each pixel to fuse N images using the weights derived from our quality assessment. To ensure consistent results, we normalize the values of N weight maps so that their sum equals 1 at each pixel location (i, j) to achieve a balanced amalgamation of the images:

$$\widehat{W}_{ij,k} = \left[\sum_{k'=1}^{N} W_{ij,k'}\right]^{-1} W_{ij,k} \tag{9}$$

The resultant image, denoted as R, can be derived through a weighted blending of the input images.

$$R_{ij} = \sum\nolimits_{k=1}^{N} \widehat{W}_{ij,k} I_{ij,k} \tag{10}$$

Let I_k represent the k^{th} input image in the sequence. We are employing Eqs. (9) and (10) adjusting the weights in various ways invariably result in noticeable seams within the composite image. This issue primarily stems from the differences in absolute intensities among the images to be merged, which are due to varying exposure times. While applying Gaussian filters to smooth the weight maps can mitigate the abrupt transitions in the weight maps, this approach often leads to undesirable halo effects around image edges. It may result in information leakage at object boundaries. To address the issue of seams, we define the l^{th} level of the Laplacian pyramid decomposition of image A as $L\{A\}^l$, and for image B, we define its Gaussian pyramid as $G\{B\}^l$. Subsequently, we blend the coefficients (pixel intensities across different pyramid levels) in a manner akin to that described by the equation. This methodological approach facilitates the seamless integration of images A and B by meticulously combining their respective pyramid representations, thereby ensuring a coherent transition between the two images without the appearance of noticeable seams.

$$L\{R\}_{ij}^{l} = \sum\nolimits_{k=1}^{N} G\{\widehat{W}\}_{ij,k}^{l} L\{I\}_{ij,k}^{l} \tag{11}$$

Thus, each level of the derived Laplacian pyramid is computed as the weighted average of the original Laplacian decomposition at the same level, with the weights being sourced from the l^{th} level of the Gaussian pyramid of the weight map. The pyramid L is then collapsed to produce the final image R. Multiresolution blending has proven to be highly effective in avoiding seams, as it merges image features rather than intensities, ensuring that the final composite image R encapsulates the most salient features from the inputs. This image-blending technique leverages the advantages of multiresolution analysis to seamlessly integrate features from different images.

Employing a Laplacian pyramid to capture detailed features at various scales and a Gaussian pyramid to determine the blending weights ensures a smooth transition between images. The final step of collapsing the

Laplacian pyramid effectively reconstructs the blended image, preserving the essential characteristics of source images while eliminating visible seams. This technique enhances the clarity and detail of the resultant image. It ensures a seamless integration of the various attributes present in the input images, leading to a coherent and visually appealing output.

Experiments and analysis

Experiments

Our experimental study centers on image processing using the Sony subset of the SID[24] dataset within a self-attention U-Net framework. To evaluate the effectiveness of our approach, we train the model using minimally conditioned data, specifically with the shortest exposure time of 0.1 s as input. The training process extends over 4000 epochs and utilizes the Adam optimizer[47], with an L1 loss function as the loss metric. Initially, the model is trained using a pre-existing framework to generate images for specific exposure settings. However, this results in overexposure in the output images. We make subsequent adjustments to the model parameters to address this issue, effectively mitigating the overexposure problem. Consequently, the model can produce images with balanced HDR results. This systematic approach underscores the potential of our self-attention U-net framework in improving image quality while skillfully managing exposure levels.

We utilize 161 sets of images for training. As illustrated in Fig. 8, our initial loss is approximately 0.18 at the start of the training process. It rapidly decreases to 0.04 after about 100 epochs. Following this, the loss continues to decline steadily, dropping from 0.03 to around 0.02 by the 1750th epoch, after which it stabilizes and gradually decreases further. The lowest value of loss is 0.019097764. After obtaining the image output from the self-attention network, we utilize the enhanced or weakened exposure network to generate a series of exposure images, as illustrated in Fig. 9. We determine which images are suitable for fusion to produce the final output by analyzing the differences between adjacent pixels.

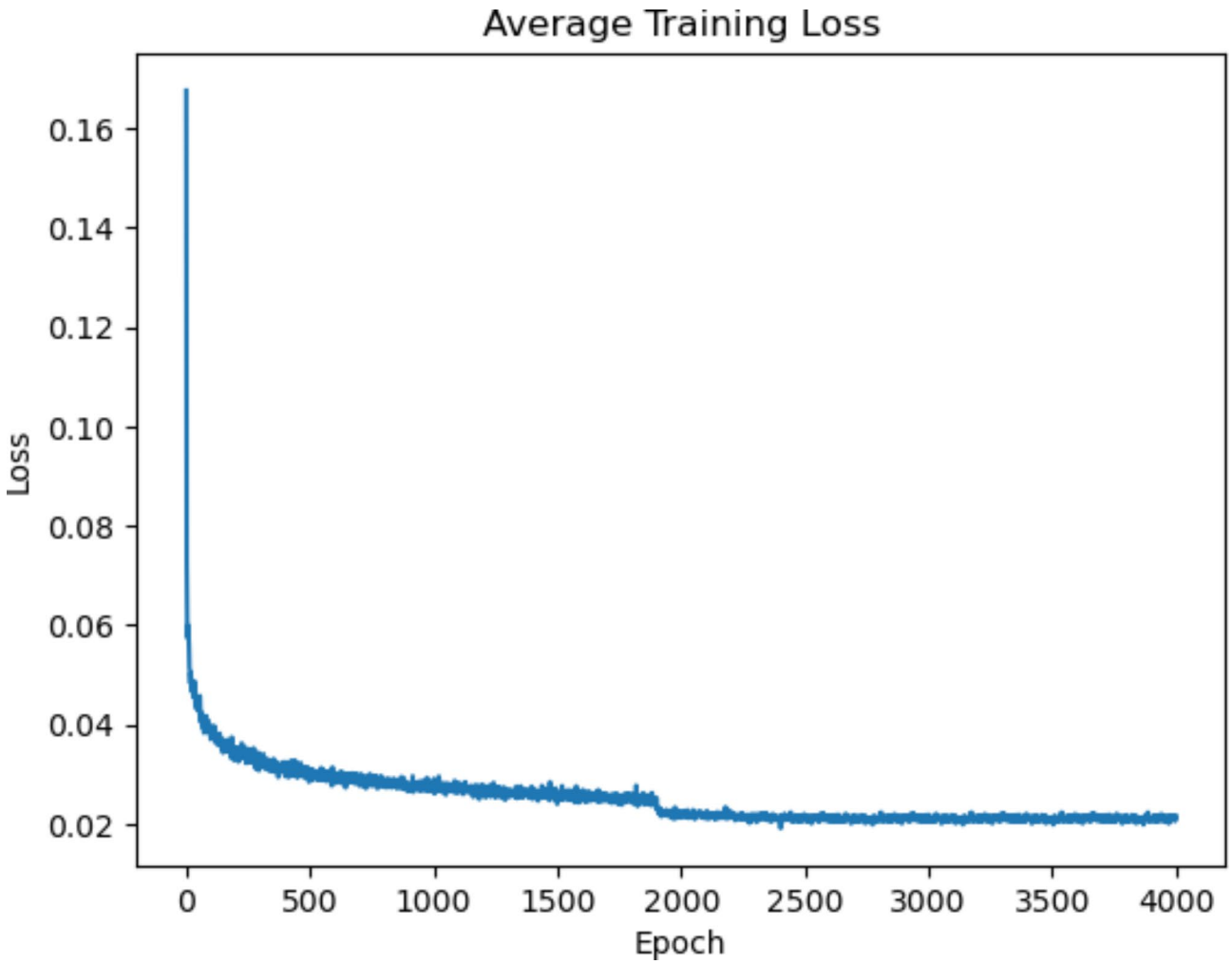

Fig. 8. Average loss over 4000 epochs. After starting training, the loss dropped from 0.18 to around 0.02 at the end.

Fig. 9. Filtering using pixel difference between the two images in a series of exposure images generated.

Ablation studies

Ablation experiments at each stage in four scenarios

An extensive comparison is conducted between the proposed method and various existing methods, including traditional methods[1,7], single image HDR image generation[20], SID[24], our single-step self-attention U-net, SID U-net[24] combined with our HDR linear exposure enhancement fusion. The comparative results are presented in Fig. 10. The figure demonstrates that the traditional methods[1,7] and single-image HDR generation[20] employed to achieve restoration on this experimental dataset are ineffective. On the other hand, models trained by the end-to-end approach effectively solve the problem. Nevertheless, unoptimized end-to-end training remains inadequate, and thus, we investigate the key steps in solving the different problems through ablation experiments.

Detailed comparison of ablation experiments

Consequently, we divide our ablation experiment into four groups, and the image results are displayed in Fig. 11. The first group uses SID U-net[24] as the baseline, the second group utilizes our single-step self-attention U-net network, the third group adopts the baseline in combination with our HDR linear exposure fusion module, and the fourth group encompasses our full pipeline (self-attention U-net + CCM module + HDR linear exposure enhancement fusion module). The specific data results can be found in Table 1, in which the red font color represents the best result, the blue color represents the second best result, and the upward arrows indicate that the higher the data, the better the image quality generated. Comparing the results of the first and second groups, the overall image shows better balance - although issues related to color casting and underexposure remain prominent (Fig. 12).

Image quality is commonly assessed using metrics such as the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM). PSNR, a prevalent metric, quantifies the pixel-level errors between images to offer an error-sensitive evaluation of image quality. However, accounting for the visual characteristics of human eyes may lead to discrepancies between the metric assessment and the subjective human perception. On the other hand, SSIM evaluates image similarity by considering key attributes such as luminance, contrast, and structural information, aligning more closely with human visual perception. Compared to the unimproved U-net network, our self-attention U-net[24] improves PSNR by 0.21dB and SSIM by 0.001 in a single step. Compared with the first and third groups, the third group of images exhibits significantly improved exposure and richer image details, albeit the color cast problem persists. The data indicators demonstrate that PSNR is increased by 1.05dB and SSIM by 0.007. Ultimately, when comparing the results of the third and fourth groups, the fourth group of images displays a more refined overall white balance, addresses the color cast issue, and presents clearer image details. The data indicators reveal that PSNR is increased by 0.77dB and SSIM by 0.005.

The self-attention U-net network effectively reduces noise in the image. Including the CCM color correction module resolves color casting issues and restores the image's white balance. The HDR reconstruction step effectively addresses underexposure in imaging, regardless of whether the self-attention U-net network is utilized. This evidence validates the effectiveness of each stage in our approach, showcasing the remarkable ability of our technique to preserve and reconstruct image structural integrity and perceptual quality under extremely low light conditions.

Comparison with other state-of-the-art methods

In our evaluation, we thoroughly analyze the proposed pipeline using the SID[24] dataset and compare the performance with literature studies. We examine both single-stage processes that utilize only self-attention

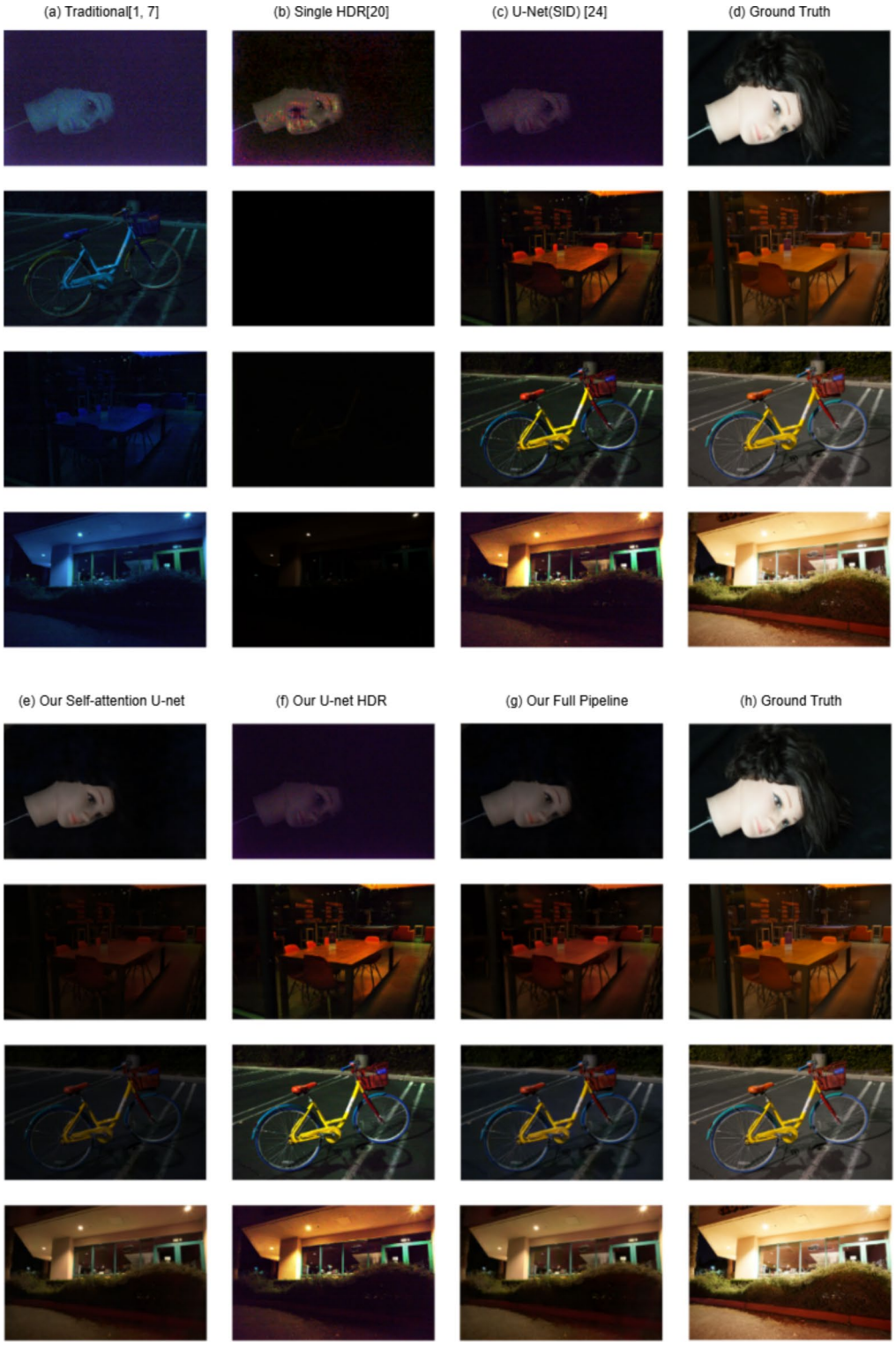

Fig. 10. Comparison of results obtained by different methods, in which Column (**a**) shows the combination of traditional mathematical methods, Column (**b**) uses a single image to generate the HDR image, Columns (**c**), (**e**), (**f**), and (**g**) are the results of end-to-end single-step and multi-step approaches, and Column (**d**) and (**h**) is the ground truth.

models, such as SID[24], DID[28], SGN[10], LLPackNet[46], RRT[47], and multi-stage processes involving the entire pipeline, such as EEMEFN[48], LDC[49], MCR[31], RRENet[32], and DNF[41].

Even when processing raw dark input data, we achieve excellent results, including the highest performance metrics, as shown in Table 2. In the initial stage, our self-attention U-Net network only addresses the denoising

Fig. 11. (**a**-**d**) Four groups of ablation experiments and zoom-in images on the text part to demonstrate the validity of experiments by the reducibility of the background and font colors, (**e**) the ground truth.

Method	PSNR ↑	SSIM ↑
Traditional[1,7]	28.54	0.355
Single -HDR[20]	28.07	0.220
U-net(SID)[24]	28.96	0.787
U-net + HDR	**30.01**	**0.794**
Self-Attention U-net	29.17	0.788
Full Pipeline (Self-Attention U-net + HDR)	**30.78**	**0.799**

Table 1. Result comparison among various methods, using PSNR and SSIM metrics.

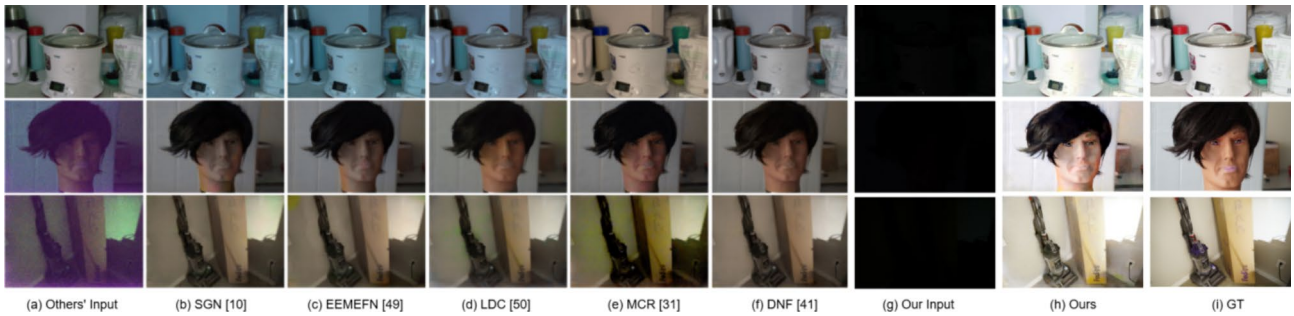

Fig. 12. Comparison between the proposed linear enhancement exposure fusion method and various learning-based methods developed in recent years.

performance. It does not tackle the color cast problem of the image, ranking as the second-best in denoising. However, SGN[10] obtains the highest score, but the color cast problem still needs to be addressed. Our subsequent steps address the color cast and exposure issues, elevating all metrics to the highest levels. Compared with the second-best results from DNF, the proposed pipeline improves PSNR by 0.16 dB and attains the highest SSIM value. Qualitatively, as demonstrated in Fig. 12, our linear enhancement exposure fusion method effectively balances exposure, corrects color bias, prevents overexposure, and preserves rich texture details throughout the image.

Conclusion

In this research, we develop a novel method for enhancing extremely low-light images. This method leverages a self-attention mechanism and a U-Net network combined with HDR reconstruction techniques to effectively reduce black-level noise and restore the true color of the images. Following color correction, the method employs a linear exposure enhancement model to generate an image with increased exposure. The algorithm selects the best images and merges them to produce the final result, enhancing the image's brightness details. Compared to traditional methods, direct HDR reconstruction, and various CNN training methods applied to images in different scenarios, our method consistently demonstrates superior PSNR and SSIM performance. This progress underscores the effectiveness of our proposed method in enhancing the quality of images captured in extremely low-light environments.

Category	Method	PSNR ↑	SSIM ↑
Single Stage	SID[24]	28.96	0.787
	DID[28]	29.16	0.785
	SGN[10]	**29.28**	**0.790**
	LLPackNet[48]	27.83	0.755
	RRT[49]	28.66	0.790
	Self-Attention U-net	**29.17**	**0.788**
Multi-Stage	EEMEFN[50]	29.60	0.795
	LDC[51]	29.56	**0.799**
	MCR[31]	29.65	**0.797**
	RRENet[32]	29.17	0.792
	DNF[41]	**30.62**	**0.797**
	Full Pipeline (Self -Attention U-net + HDR)	**30.78**	**0.799**

Table 2. Comparison of ablation studies using PSNR and SSIM metrics.

It should be noted that the proposed approach has limitations; thus, developing the mitigation solutions will extend our research in the future. First, the multi-step approach requires a long computational time to process a high-resolution 4D RGGB image, and the model's training consumes extensive resources, too. In the future, more efficient algorithms must be developed to address the issue. Second, the pre-trained model generates a continuous exposure image that exceeds the exposure of the ground truth and is more mathematically computed at the time of fusion. Therefore, the darker areas of the image are optimized for high exposure during the synthesis, such as the face of the second portrait image in Fig. 12. In the future, a second stage of end-to-end training can be explored to target the exposure of the image for better results. Third, this work can be further improved in the image fusion stage and more evaluation indicators could be adopted to further validate the effectiveness of our method. Lastly, it is recognized that a feedback mechanism would present a valuable opportunity to enhance the proposed pipeline. Implementing such a mechanism in training will be actively explored in the future.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 6 November 2024; Accepted: 20 January 2025
Published online: 22 January 2025

References

1. Hu, Z., Cho, S., Wang, J. & Yang, M. H. Deblurring low-light images with light streaks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3382–3389 (2014). https://doi.org/10.1109/CVPR.2014.432.
2. Rudin, L. I., Osher, S. & Fatemi, E. Nonlinear total variation based noise removal algorithms. *Phys. D: Nonlinear Phenom.* **60**(1–4), 259–268. https://doi.org/10.1016/0167-2789(92)90242-F (1992).
3. Portilla, J., Strela, V., Wainwright, M. J. & Simoncelli, E. P. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Trans. Image Process.* **12**(11), 1338–1351. https://doi.org/10.1109/TIP.2003.818640 (2003).
4. Vaswani, A. et al. Attention is all you need. *Adv. Neural. Inf. Process. Syst.* **30**. https://doi.org/10.7551/mitpress/7503.003.0158 (2017).
5. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18* 234–241 (Springer, 2015). https://doi.org/10.1007/978-3-319-24574-4_28
6. Liu, Y. C., Chan, W. H. & Chen, Y. Q. Automatic white balance for digital still camera. *IEEE Trans. Consum. Electron.* **41**(3), 460–466. https://doi.org/10.1109/30.468045 (1995).
7. Rahman, S., Rahman, M. M., Abdullah-Al-Wadud, M., Al-Quaderi, G. D. & Shoyaib, M. An adaptive gamma correction for image enhancement. *EURASIP J. Image Video Process.* **2016**(1), 1–11. https://doi.org/10.1186/s13640-016-0138-1 (2016).
8. Jiang, H. et al. Learning the image processing pipeline. *IEEE Trans. Image Process.* **26**(10), 5032–5042. https://doi.org/10.1109/TIP.2018.2872858 (2017).
9. Mairal, J., Bach, F., Ponce, J., Sapiro, G. & Zisserman, A. Non-local sparse models for image restoration. In *2009 IEEE 12th International Conference on Computer Vision. IEEE* 2272–2279 (2009). https://doi.org/10.1109/ICCV.2009.5459452.
10. Gu, S. et al. Self-guided network for fast image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2511–2520 (2019). https://doi.org/10.1109/ICCV.2019.00260.
11. Liu, Z. et al. Fast burst images denoising. *ACM Trans. Graphics (TOG)* **33**(6), 1–9. https://doi.org/10.1145/2661229.2661277 (2014).
12. Godard, C., Matzen, K. & Uyttendaele, M. Deep burst denoising. In *Proceedings of the European Conference on Computer Vision (ECCV)* 538–554 (2018). https://doi.org/10.1007/978-3-030-01267-0_33.
13. Mildenhall, B. et al. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2502–2510 (2018). https://doi.org/10.1109/CVPR.2018.00265.
14. Hasinoff, S. W. et al. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Trans. Graphics (ToG)* **35**(6), 1–12. https://doi.org/10.1145/2980179.2980254 (2016).
15. Yan, Q. et al. Multi-scale dense networks for deep high dynamic range imaging. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE* 41–50 (2019). https://doi.org/10.1109/WACV.2019.00012.
16. Wu, S. et al. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)* 117–132 (2018). https://doi.org/10.1109/ICCV.2001.937492.

17. Kalantari, N. K. & Ramamoorthi, R. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graphics* **36**(4), 144:1-144:12. https://doi.org/10.1145/3072959.3073609 (2017).
18. Liu, Y. L. et al. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1651–1660 (2020). https://doi.org/10.1109/CVPR42600.2020.00172.
19. Lee, S., An, G. H. & Kang, S. J. Deep recursive HDRI: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)* 596–611 (2018). https://doi.org/10.1007/978-3-030-01216-8_37.
20. Endo, Y., Kanamori, Y. & Mitani, J. Deep reverse tone mapping. *ACM Trans. Graphics* **36**(6), 1771–17711. https://doi.org/10.1145/3130800.3130834 (2017).
21. Kovaleski, R. P. & Oliveira, M. M. High-quality reverse tone mapping for a wide range of exposures. In *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images. IEEE* 49–56 (2014). https://doi.org/10.1109/SIBGRAPI.2014.29.
22. Huo, Y., Yang, F., Dong, L. & Brost, V. Physiological inverse tone mapping based on retina response. *Visual Comput.* **30**(5–7), 507–517. https://doi.org/10.1007/s00371-013-0875-4 (2014).
23. Le, P. H. et al. Single-image HDR reconstruction by multi-exposure generation,. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* 4063–4072 (2023). https://doi.org/10.1109/WACV56688.2023.00405.
24. Chen, C. et al. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3291–3300 (2018). https://doi.org/10.1109/CVPR.2018.00347.
25. Chen, C., Chen, Q., Do, M. N. & Koltun, V. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 3185–3194 (2019). https://doi.org/10.1109/ICCV.2019.00328.
26. Cai, Y. & Kintak, U. Low-light image enhancement based on modified U-Net. In *2019 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR). IEEE* 1–7 (2019). https://doi.org/10.1109/ICWAPR48189.2019.8946456.
27. Schwartz, E., Giryes, R. & Bronstein, A. M. DeepISP: Toward learning an end-to-end image processing pipeline. *IEEE Trans. Image Process.* **28**(2), 912–923. https://doi.org/10.1109/TIP.2018.2872858 (2018).
28. Maharjan, P. et al. Improving extreme low-light image denoising via residual learning. In *2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE* 916–921 (2019). https://doi.org/10.1109/ICME.2019.00162.
29. Zamir, S. W. et al. Learning digital camera pipeline for extreme low-light imaging. *Neurocomputing* **452**, 37–47. https://doi.org/10.1016/j.neucom.2021.04.076 (2021).
30. Cao, Y. et al. Physics-guided iso-dependent sensor noise modeling for extreme low-light photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5744–5753 (2023). https://doi.org/10.1109/CVPR52729.2023.00556.
31. Dong, X. et al. Abandoning the Bayer-Filter to See in the Dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 17431–17440 (2022). https://doi.org/10.1109/CVPR52688.2022.01691.
32. Jin, X. et al. DNF: Decouple and feedback network for seeing in the dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 18135–18144 (2023). https://doi.org/10.1109/CVPR52729.2023.01739.
33. Bhandari, A. K., Subramani, B. & Veluchamy, M. Multi-exposure optimized contrast and brightness balance color image enhancement. *Digit. Signal Proc.* **123**, 103406. https://doi.org/10.1016/j.dsp.2022.103406 (2022).
34. Subramani, B., Bhandari, A. K. & Veluchamy, M. Optimal Bezier curve modification function for contrast degraded images. *IEEE Trans. Instrum. Meas.* **70**, 1–10. https://doi.org/10.1109/TIM.2021.3073320 (2021).
35. Subramani, B. & Veluchamy, M. Cuckoo search optimization-based image color and detail enhancement for contrast distorted images. *Color. Res. Appl.* **47**(4), 1005–1022. https://doi.org/10.1007/s11042-020-08870-1 (2022).
36. Veluchamy, M. & Subramani, B. Fuzzy dissimilarity contextual intensity transformation with gamma correction for color image enhancement. *Multimed. Tools Appl.* **79**(27), 19945–19961. https://doi.org/10.1007/s11042-020-08870-1 (2020).
37. Anaya, J. & Barbu, A. RENOIR—A dataset for real low-light image noise reduction. *J. Vis. Commun. Image Represent.* **51**, 144–154. https://doi.org/10.1016/j.jvcir.2018.01.012 (2018).
38. Plotz, T. & Roth, S. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1586–1595 (2017). https://doi.org/10.1109/CVPR.2017.294.
39. Nam, S., Hwang, Y., Matsushita, Y. & Kim, S. J. A holistic approach to cross-channel image noise modeling and its application to image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1683–1691 (2016). https://doi.org/10.1109/CVPR.2016.186.
40. Xu, J. et al. Real-world noisy image denoising: A new benchmark. arXiv preprint arXiv:1804.02603 (2018). https://doi.org/10.1109/TIP.2018.2811546.
41. Huang, H. et al. Towards low light enhancement with raw images. *IEEE Trans. Image Process.* **31**, 1391–1405. https://doi.org/10.1109/TIP.2022.3140610 (2022).
42. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473 (2014). https://doi.org/10.18653/v1/2021.emnlp-main.1.
43. Luong, M. T., Pham, H. & Manning, C. D. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015). https://doi.org/10.18653/v1/D15-1166.
44. Oktay, O. et al. Attention U-Net: Learning where to look for the pancreas. arXiv Preprint arXiv:1804.03999 (2018). https://doi.org/10.1161/STROKEAHA.110.591214.
45. Grossberg, M. D. & Nayar, S. K. What is the space of camera response functions? In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. IEEE, 2: II-602* (2003). https://doi.org/10.1109/CVPR.2003.1211522.
46. Mertens, T., Kautz, J. & Van Reeth, F. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications (PG'07). IEEE* 382–390 (2007). https://doi.org/10.1109/PG.2007.17.
47. P Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv Preprint arXiv:1412.6980 (2014). https://doi.org/10.1080/10556788.2014.891592.
48. Lamba, M., Balaji, A. & Mitra, K. Towards fast and light-weight restoration of dark images. arXiv Preprint arXiv:2011.14133 (2020). https://doi.org/10.1109/WACV56688.2023.00489.
49. Lamba, M. & Mitra, K. Restoring extremely dark images in real time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3487–3497 (2021). https://doi.org/10.1109/CVPR46437.2021.00349.
50. Zhu, M. et al. EEMEFN: Low-light image enhancement via edge-enhanced multi-exposure fusion network. In *Proceedings of the AAAI Conference on Artificial Intelligence* 13106–13113 (Vol. 34, No. 07) (2020). https://doi.org/10.1609/aaai.v34i07.7013.
51. Xu, K. et al. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2281–2290 (2020). https://doi.org/10.1109/CVPR42600.2020.00235.

Author contributions

Y. H.: Writing-Reviewing, Editing, Conceptualization, Methodology, and Software. X. Z.: Validation, Supervision, Editing and Project Administration.F. Y.: Methodology, Editing, and Rewriting.J. S.: Supervision, Editing, and Rewriting.K. U: Investigation, Methodology, Project Administration.J. F.: Conceptualization and MethodologyY. P.: Software and Validation.C. D.: Validation. All authors have reviewed the manuscript and are in agreement for submission for peer review.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to J.S. or U.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.