Neural Networks 162 (2023) 557–570

Contents lists available at ScienceDirect

Neural Networks

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

Restoration and enhancement on low exposure raw images by joint demosaicing and denoising

Jiaqi Ma [a], Guoli Wang [c], Lefei Zhang [a,b,*], Qian Zhang [c]

[a] Institute of Artificial Intelligence and School of Computer Science, Wuhan University, Wuhan, 430072, PR China
[b] Hubei Luojia Laboratory, Wuhan 430072, PR China
[c] Horizon Robotics, Beijing, 100089, PR China

ARTICLE INFO

Article history:
Received 23 September 2022
Received in revised form 31 January 2023
Accepted 12 March 2023
Available online 15 March 2023

Keywords:
Raw images
Joint demosaicing and denoising
Image enhancement

ABSTRACT

Restoring high quality images from raw data in low light is challenging due to various noises caused by limited photon count and complicated Image Signal Process (ISP). Although several restoration and enhancement approaches are proposed, they may fail in extreme conditions, such as imaging short exposure raw data. The first path-breaking attempt is to utilize the connection between a pair of short and long exposure raw data and outputs RGB images as the final results. However, the whole pipeline still suffers from some blurs and color distortion. To overcome those difficulties, we propose an end-to-end network that contains two effective subnets to joint demosaic and denoise low exposure raw images. While traditional ISP are difficult to image them in acceptable conditions, the short exposure raw images can be better restored and enhanced by our model. For denoising, the proposed Short2Long raw restoration subnet outputs pseudo long exposure raw data with little noisy points. Then for demosaicing, the proposed Color consistent RGB enhancement subnet generates corresponding RGB images with the desired attributes: sharpness, color vividness, good contrast and little noise. By training the network in an end-to-end manner, our method avoids additional tuning by experts. We conduct experiments to reveal good results on three raw data datasets. We also illustrate the effectiveness of each module and the well generalization ability of this model.

1. Introduction

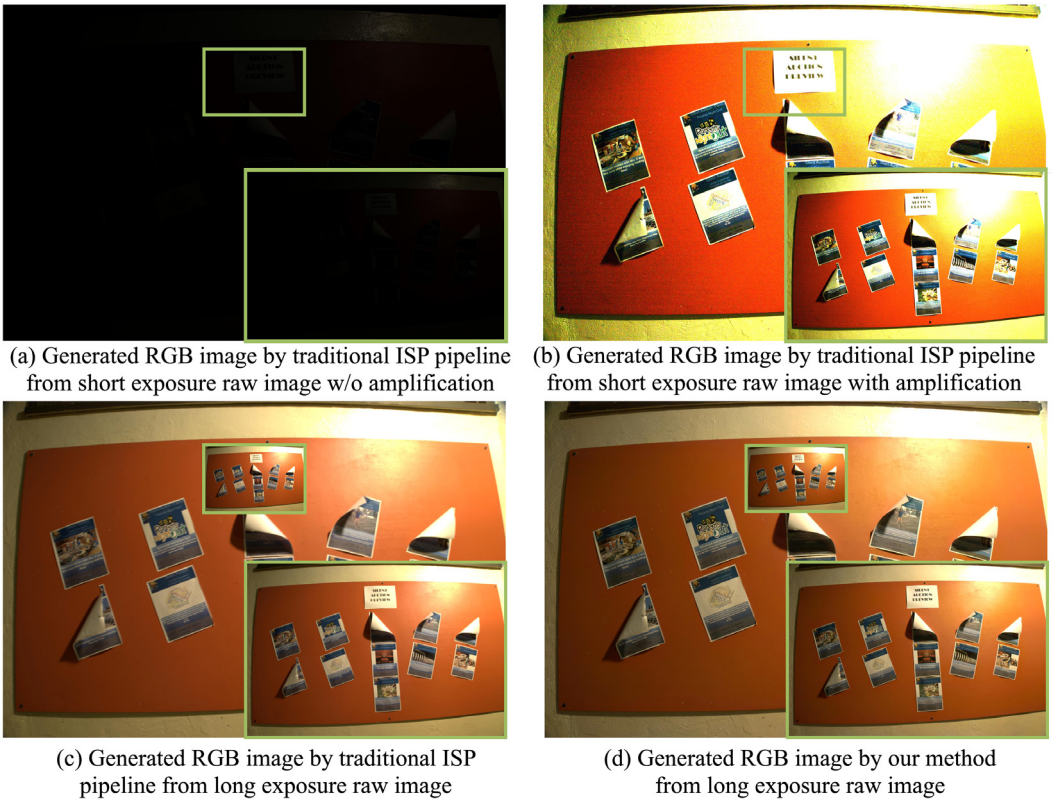
Raw images refer to unprocessed data from the image sensor and are mainly captured by CMOS in cameras. Each CMOS is sensitized by light and the camera receives corresponding digital signals to generate a raw image (Fukushima et al., 1983). Although color filter arrays differs to manufacturers, all of them place a single color (such as Red, Green and Blue) CMOS on a pixel location. It means that we cannot directly see visible contents from a raw image. Through Image Signal Processing (ISP) pipelines containing demosaicing and denoising, those raw images can be turned into RGB images. Nevertheless, the visual results for short exposure raw data are unacceptable like Fig. 1(a), so a simple method for better view is to amplify every pixel with a certain ratio and get noisy results, such as Fig. 1(b). In this paper, we try to device an effective method for restoring and enhancing short exposure raw data to make up the huge differences between short and long raw data and generate satisfying results like long exposure raw images.

Traditional ISP is manually adjusted by experts and may fail in specific conditions such as low light and short exposure. For instance, Fig. 1(a) is a RGB format image processed by traditional ISP modules from a low exposure raw image, and the contents are nearly invisible. Even amplifying the input with a ratio, the output in Fig. 1(b) still contains color distortion and many noisy points in details. On contrast, a long exposure raw image can be processed relatively well by traditional ISP modules, like Fig. 1(c). A simple way to relieve this dilemma is to increase the brightness, but the existing noise and mosaic will be amplified and even cause an obvious color shift. Some researchers try collaborative filters (Dabov, Foi, Katkovnik, & Egiazarian, 2007), noise embedding (Chatterjee, Joshi, Kang, & Matsushita, 2011) and burst images (Liu, Yuan, Tang, Uyttendaele, & Sun, 2014), but results are still imperfect. Based on this observation, we consider that there may be a better way to photographing in a short time or in a low ambient light condition (Those two situations can be concluded as low exposure time).

To overcome above problems, deep learning methods, such as Convolutional Nerual Networks (CNNs), are applied to low exposure raw images. To be specific, CNNs are utilized to simulate several operations (white balance, demosaicing, denoising, color correction, tone mapping, sharpening, etc.) of traditional ISP modules. Many attempts have been done, such as CameraNet (Liang,

* Corresponding author at: Institute of Artificial Intelligence and School of Computer Science, Wuhan University, Wuhan, 430072, PR China.
E-mail address: zhanglefei@whu.edu.cn (L. Zhang).

https://doi.org/10.1016/j.neunet.2023.03.018
0893-6080/



(a) Generated RGB image by traditional ISP pipeline from short exposure raw image w/o amplification

(b) Generated RGB image by traditional ISP pipeline from short exposure raw image with amplification

(c) Generated RGB image by traditional ISP pipeline from long exposure raw image

(d) Generated RGB image by our method from long exposure raw image

Fig. 1. The illustration of a short exposure raw image processed by the traditional ISP pipeline w/o and with amplification, corresponding long exposure one processed by the traditional ISP pipeline, and generated RGB image by our method from Sony dataset.

Cai, Cao, & Zhang, 2021), DeepISP (Schwartz, Giryes, & Bronstein, 2019) and PyNet (Ignatov, Gool, & Timofte, 2020). However, it is hard to design separable modules for every ISP operations due to the lack of handcrafted priors and the vulnerable coordinated states in a low exposure condition. Instead, researchers find that the key components for reconstruction are denoising and demosaicing (Chatterjee et al., 2011; Hirakawa & Parks, 2006; Ma, Yan, Zhang, Wang, & Zhang, 2022; Xing & Egiazarian, 2021). A scheme called Joint Denoising and Demosaicing (JDD) is proposed. Noises can be removed from raw data by denoising, and RGB images can be reconstructed by demosaicing. With well designed CNN models under the JDD scheme, we can get high quality RGB results from low quality raw data and even tune them for better visual effects than traditional ISP pipelines.

The motivations behind this work are mainly twofold. On one hand, in order to restore raw data of high quality, the essence of raw image learning is denoising. Recently, researchers have tried a lot on learning-based raw data denoising. Brooks et al. (2019) simulate the ISP pipeline to denoise but fail in learning from different domains. From the perspective of perceptual, low exposure raw images still have different distribution compared with long exposure ones. Hence, results may turn into another color tone and perform bad on non-referenced perception metrics. Although Liang, Guo, Gu, Zhang, and Zhang (2020) use burst sequences, it is restricted by the limited burst data. Wei, Fu, Yang, and Huang (2020) also design a noise model based on the characteristics of CMOS and show good results. Ren et al. (2022) design image restoration tasks in manner of low-rank representation and reach good performance. Zhang, Song, Du, and Zhang (2021) also propose a novel view for visual data completion. Nevertheless, the additional images from different domains make it hard to be realized in real world. Hence, our goal is to perform raw denoising well with a concise structure and limited data. Inspired by Wang, Huang et al. (2020), we confirm the practical of U-shaped structure for raw denoising and propose a Short2Long raw restoration subnet to remove noisy points and maintain the useful recovery information from under-exposed raw images.

On the other hand, the generated images still have flaws like color distortion and removed details, resulting in low structural metrics like Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM). Chen, Chen, Xu, and Koltun (2018) firstly illustrate the feasibility of CNNs to replace the traditional ISP pipeline, but the proposed method results in artifacts and blurs on the generated low exposure RGB images. Maharjan et al. (2019) relieve those phenomena by residual connections, but only show similar results as SID. Xu, Yang, Yin, and Lau (2020) focus on the decomposition of low and high frequency information, while the details of generated images are missing. The aforementioned observation encourages us to design a demosaicing subnet related to color correction — the Color consistent RGB enhancement subnet. We focus on how to rearrange those single channel data to be close as the details and color tone of the original ones.

In this paper, we propose a joint demosaicing and denoising architecture containing two effective subnets that is capable of generating pseudo well-exposed raw images and corresponding RGB images with the desired attributes: sharpness, color vividness, good contrast and little noise. Fig. 1(d) shows the image produced by the proposed approach, and its details and visual representations are much better than traditional ISP. Our end-to-end network takes the raw data captured in extreme low light as input and generates corresponding both well-exposed raw and RGB images. The Short2Long raw restoration subnet restrains noisy points on original raw data and provides high quality raw data for the subsequent network. The Color consistent RGB enhancement subnet receives pseudo well exposed data. By combining Pixel-wise reconstruction loss $\mathcal{L}_{rec}$ and Perceptual similarity loss $\mathcal{L}_{per}$, our method realizes the joint demosaicing

and denoising framework and learns the network in a supervised manner.

Overall, we summarize the contributions of this paper as follows:

- We propose a Short2Long raw restoration strategy to process low quality raw images and restore suitable pseudo raw ones for the subsequent subnet. It extracts sufficient information and removes enough noisy points with the concise U-shaped structure, thus mainly focuses on the denoising task.
- We device a Color consistent RGB enhancement mechanism to rearrange the raw data and generate the final RGB outputs. With the residual connection and channel-wise attention, this branch reconstructs the color spatial information and emphasizes more on the demosaicing task.
- Both quantitative and qualitative experiments are conducted on several datasets to demonstrate the superiority of our proposed method. Ablation studies prove the effectiveness of every modules. By applying the pre-trained model on low quality raw images from smartphones, we also illustrate its promising generalization ability on unseen raw images.

2. Related work

2.1. CNN-based ISP

Traditional ISP methods are designed and tuned by experienced experts. Those algorithms vary depending on brands of camera sensors and have parameters which are estimated by visual effect. With the development of deep learning, researchers make many attempts on integrating those learning based approach to traditional ISP pipeline. Jiang, Tian, Farrell, and Wandell (2017) design a framework to find the affine mapping. This method cluster the raw image patches based on simple features and map the raw patches to the sRGB patches, but its simple parametric model results in limited performance. Chen et al. (2018) propose a novel U-net architecture for restoring and enhancing low light raw data. They present a completely new raw dataset for this task and show a promising future. The noisy and dimly short exposure raw images can be converted to clear and colorful RGB images. Schwartz et al. (2019) prefer to replace the whole ISP pipeline with an end-to-end CNN architecture called DeepISP. Morawski et al. (2022) illustrate a minimal neural ISP, which can improve the ability of machine cognition in low-light conditions. Liang et al. (2021) extend the pipeline to two stages and reach a better performance. Liu, Lai et al. (2020) device a reverse ISP pipeline and learn to reconstruct a high dynamic range image from single input. From images to videos, Jiang and Zheng (2019) propose a low light video enhancement method. Generally speaking, those existing methods only focus on some parts of ISP pipeline and still suffer problems such as bad performance in low light conditions.

2.2. Joint demosaicing and denoising

Image demosaicing and image denoising are two classic low-level CV tasks and are strongly correlated with each other. Demosaicing aims to restore raw data to visualized RGB images (It is considered as incomplete data caused by the design of CFA in sensors.). Nevertheless, the raw data are easily contaminated with noises, so the final results after direct demosaicing are unsatisfying. Naturally, researchers try to perform demosaicing and denoising simultaneously. Joint demosaicing and denoising can relieve the error accumulation of a single process (Liu, Jia & Tian, 2020).

JDD methods can be divided into optimization-based methods and learning-based methods. The traditional methods are based on some heuristics, such as Condat and Mosaddegh (2012), Klatzer, Hammernik, Knöbelreiter, and Pock (2016) and Tan, Zeng, Lai, Liu, and Zhang (2017). Recently, deep-learning methods have outperformed traditional ones both in quantitative and quality evaluations. Gharbi, Chaurasia, Paris, and Durand (2016) train a CNN network on millions of images, which achieves state-of-the-art performance. Liu, Jia and Tian (2020) propose a self guidance based method supervised by density-map and green channel. Xing and Egiazarian (2021) design a triple joint learning framework including demosaicing, denoising and super-resolution. Apart from these supervised approaches, Ehret, Davy, Arias, and Facciolo (2019) focus on over-fitting of raw data bursts in an unsupervised manner. Zhang et al. (2022) device a novel color consistent network, which jointly keeps color information and enhances the illumination. The trend of JDD motivates us to explore better solutions to restore and enhance low exposure raw data.

2.3. Low exposure image enhancement

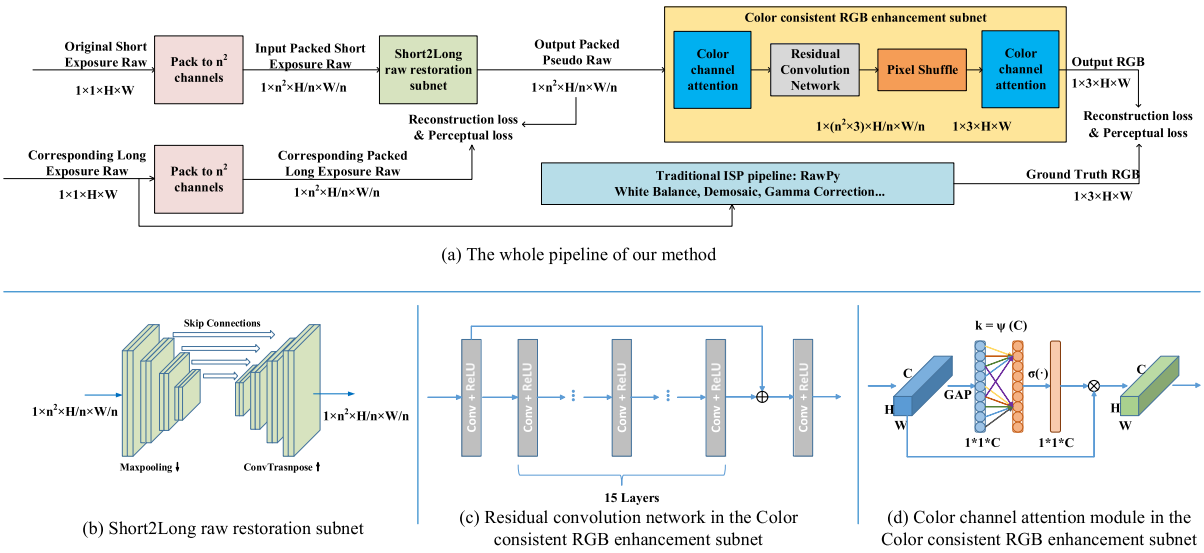
Low exposure images have been discussed for a long time, and both traditional and deep learning based methods have been applied to recover human-being welcoming and high quality images. For those traditional techniques, Stark (2000) utilize histogram equalization to simulate the White Balance process of the full size image. Besides, enlightened by human vision system, Farid (2001) discuss Gamma Correction and enhances low light images by increasing the brightness of dark regions while compressing bright pixels. Other methods such as inverse dark channel prior (Dong et al., 2011; Malm et al., 2007), wavelet transform (Loza, Bull, & Achim, 2010), Retinex model (Park, Yu, Moon, Ko, & Paik, 2017) and illumination map estimation (Guo, Li, & Ling, 2017) also show effectiveness. Recently, deep learning based methods (Xu, Yang et al., 2020; Yang, Wang, Fang, Wang, & Liu, 2020) have surpassed those traditional ways. Lamba and Mitra (2022) design a three-stage network for extremely dark images, which is fast and accurate. Dong et al. (2022) consider the low-light image enhancement problem as combining colored raw and monochrome images. Liu, Ma, Ma, Fan, and Luo (2023) propose a general learning framework based on architecture search, which can reduce computational burden.

However, the aforementioned methods are based on the assumption that all the input images are in the standard RGB color space and already contain a good representation of ground truth. For low exposure raw data, it is affected by severe noise and mosaic, and the lack of enough illumination can cause huge color distortion and chromatic aberration. Although methods such as Tian et al. (2022), Xu, Zhang and Zhang (2020) and Tian et al. (2020) can enhance those degraded images, most of existing pipelines cannot handle those flaws correctly.

3. Proposed method

3.1. Pipeline

The whole pipeline in Fig. 2 contains two subnets with several pre-processing steps. It is trained in a supervised manner by pairs of low exposure data and corresponding long exposure ones. For testing, the model is fed with the low exposure input and generates the final RGB output. In this section, we will focus on the architectures and functions of two subnets.



(a) The whole pipeline of our method

(b) Short2Long raw restoration subnet

(c) Residual convolution network in the Color consistent RGB enhancement subnet

(d) Color channel attention module in the Color consistent RGB enhancement subnet

Fig. 2. The structure of our framework. Given a raw image of 1 channel as input captured in extremely low exposure conditions, it is packed to 4 channels firstly. The Short2Long raw restoration subnet learns corresponding pseudo long exposure raw data and the Color consistent RGB enhancement subnet uses it to generate a well-exposed sRGB image. The reconstruction loss and perceptual loss force the two subnets to produce outputs as pixel-wise and perceptually similar as possible to the reference long exposure raw and RGB image.

3.1.1. Short2Long raw restoration subnet

The subnet inherits a U-net encoder–decoder structure (Ronneberger, Fischer, & Brox, 2015) with symmetric skip connections between layers of the encoder and corresponding layers of the decoder. Our intention is to maintain the useful recovery information from short exposure raw images at most and remove pixel-level noises caused by insufficient light exposure simultaneously. The benefits of such a design for the restoration subnet are three-fold:

- It has superior performance on image restoration (Chen et al., 2018), and the concise design is easy for extension.
- It can process any size of images even full-resolution ones (i.e., raw images with 6000 × 4000 resolutions) due to its fully convolutional design.
- The skip connections between the encoder and decoder modules enable adequate propagation of context information in all levels and preserve high-resolution details.

Compared with original U-shape structure which is designed for the image segmentation task, our Short2Long subnet is more suitable for image restoration tasks. Besides multi-scale feature maps. we refer Chen et al. (2018) and propose a U-net based network for restoring and enhancing low light raw data. Unlike (Chen et al., 2018) which only modifies the channel number of encoder and decoder, our subnet utilize transpose convolutions and fuse multi-stage features to learn the structural relationship between pixels well. Also, both input and output are in manner of packed images before encoder and after decoder, so that the subnet can process bigger images which is essential to raw images of high resolution.

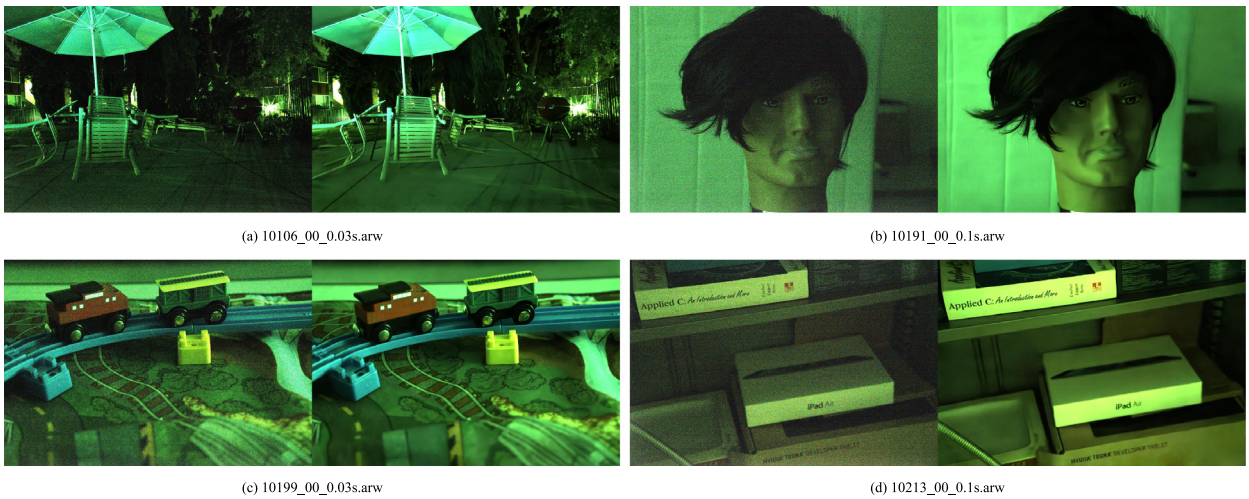
As is shown in Table 1, the Short2Long raw restoration subnet consists of Conv blocks, Transpose Conv blocks and pooling layers. Each 3 × 3 convolutional layer is followed by a Leaky ReLU nonlinearity unit. For the encoder module, max pooling operators are used for down-sampling. The decoder module also consists of 2 × 2 transpose convolutional layers for up-sampling. The skip connections is applied after every up-samplings. Finally, by appending a 1 × 1 convolutional layer, the subnet outputs a packed n^2 channel tensor with size of $\frac{H}{n}$ and $\frac{W}{n}$, which can be considered as a pseudo well-exposed raw image after rearrangement.

This subnet operates on raw images rather than RGB images, since our initial idea is to denoise those low quality data by making a connection between short and long exposure raw images. To give a clear and visual comparison to show the significance of it, we illustrate examples of inputs and outputs of the Short2Long raw restoration subnet in Fig. 3. The four pairs are short exposure raw images as inputs in left columns and corresponding pseudo long exposure raw images as outputs in right columns, respectively. All of them are one channel images which are interpolated bi-linearly and amplified by certain ratios with no more ISP processes. From Fig. 3, we can observe that our subnet removes noises in short exposure raw images well. Note that raw images are arranged by array like RGGB in single channel, simple bi-linearly interpolation results in bias on green when data is dominant by the green channel. Visualized results show that this subnet still restores fine-grained details and accurate colors in the generated pseudo long exposure raw images.

3.1.2. Color consistent RGB enhancement subnet

This subnet aims at making most use of the original color information in the reconstructed pseudo raw data from each packed color channel and rearranging those pixels to RGB images. Considering that raw images are all 1 channel, we cannot directly get Red, Green and Blue channels from single channel. A simple demosaicing operation is to copy the pixel values of defined color positions to a specific color channel, and then collect surrounding neighbor pixels to estimate missing parts, like bi-linear interpolation. In this paper, we combine the demosaicing operation into neural network, and let the network learn to demosaic. Through this subnet, raw data from the previous step performs demosaicing and is enhanced as like well-exposed images. In other words, we design this subnet to simulate several ISP steps, *i.e.*, demosaicing, white balance, gamma correction and color correction matrix (CCM). In Fig. 2(b), the Color consistent RGB enhancement subnet contains one residual convolution module, two color channel attention modules (CCA) and one pixel shuffle module.

Considering that pseudo raw data still contains some noisy points and CCA module only rearrange color channels, we need a concise module to simulate the process of denoising under supervision of high quality RGB images. Inspired by Zhang, Zuo,



(a) 10106_00_0.03s.arw (b) 10191_00_0.1s.arw

(c) 10199_00_0.03s.arw (d) 10213_00_0.1s.arw

Fig. 3. There are four example of short exposure input and their corresponding out (In each set, the left one is short exposed and the right one is pseudo long exposed.). All of raw images have only 1 channel and are interpolated bi-linearly and amplified by certain ratios with no more ISP processes. It is clear that noises in short raw images are removed well.

Table 1
The architecture of the Short2Long raw restoration subnet.

S2L subnet	Input	Operation	Out Channel	Stride
Encoder	512 × 512	Conv3 × 3 Block	32	1
	512 × 512	Conv3 × 3 Block	32	1
	512 × 512	Max Pooling	32	2
	256 × 256	Conv3 × 3 Block	64	1
	256 × 256	Conv3 × 3 Block	64	1
	256 × 256	Max Pooling	64	2
	128 × 128	Conv3 × 3 Block	128	1
	128 × 128	Conv3 × 3 Block	128	1
	128 × 128	Max Pooling	128	2
	64 × 64	Conv3 × 3 Block	256	1
	64 × 64	Conv3 × 3 Block	256	1
	64 × 64	Max Pooling	256	2
	32 × 32	Conv3 × 3 Block	512	1
	32 × 32	Conv3 × 3 Block	512	1
Decoder	32 × 32	Transpose Conv2 × 2 Block	256	2
	64 × 64	Conv3 × 3 Block	256	1
	64 × 64	Conv3 × 3 Block	256	1
	64 × 64	Transpose Conv2 × 2 Block	128	2
	128 × 128	Conv3 × 3 Block	128	1
	128 × 128	Conv3 × 3 Block	128	1
	128 × 128	Transpose Conv2 × 2 Block	64	2
	256 × 256	Conv3 × 3 Block	64	1
	256 × 256	Conv3 × 3 Block	64	1
	256 × 256	Transpose Conv2 × 2 Block	32	2
	512 × 512	Conv3 × 3 Block	32	1
	512 × 512	Conv3 × 3 Block	32	1
	512 × 512	Conv1 × 1 Block	4	1

Chen, Meng, and Zhang (2017), the residual convolution module has 17 convolutional layers with a 3 × 3 kernel which contributes to noise learning and convergence speed. Compared with simple residual convolution blocks, we make residual connections between every 3 blocks except CCA modules. Also, we still keep the longest connection from the first block to the last one. By this optimized residual connection, differences between input and output are learnt by stages, and the convergence speed can be faster, which means less training time. To accelerate the process of training and learn the color residual from each channel, there is a residual connection from the output of first layer to the input of last layer. Considering that both white balance and gamma correction are three numerical values in meta information of raw data, we do not design several specific network modules to simulate the learning process of them. Three numerical values (white balance needs two to represent ratios between channels, and gamma correction needs one to balance black and white.) can be learnt implicit by hidden layers in our subnet.

In addition, according to Wang, Wu et al. (2020), we extend the efficient channel attention as a CCA module. The insight of this CCA module is to capture the connection and consistent between each color channel. Fig. 2(c) shows details of this CCA module. Before applying this module, the C × H × W input tensor X represents C channel raw or RGB data. After global average pooling (GAP), it is transformed to the 1 × 1 × C feature map. With this feature map, we compute the adaptive kernel sizes $k = \psi(C)$. The adaptive kernel sizes k can be applied to 1-dimension convolution, and we get the rearranged weights of each color channel W. CCA module conducts element-wise product between original input X and color channel weights W, and then outputs the final C×H×W tensor X_{att}.

Compared with channel attention, the insight of this CCA module focuses more on how to capture the connection and

consistent between each color channel. Unlike channel attention which are commonly used after several convolution blocks, we utilize the CCA module before Residual convolution network and after Pixel shuffle. There, the first CCA module tends to learn the relationship of pseudo long exposure raw images. Although those raw images are output as 4 channels, we view them as packed ones and need to arrange their color orders for other procedures. For the later CCA module after Pixel shuffle, we already view those tensors as batches of RGB images. Through the last CCA, the network can easily split Red/Green/Blue color channel.

Besides, the CCM is actually a 3×4 matrix which contains weight with bias for each color channel. However, designing a module to parameterize and learning a suitable CCM results in failures by our experiments. Hence, inspired by super resolution tasks, CCM is equally replaced by a sub-pixel layer (Shi et al., 2016), which is defined as the Pixel_shuffle Class in PyTorch. It is applied at the bottom of our subnet similar with traditional ISP pipelines and can remap the channels to obtain RGB images with the same spatial resolution as its original raw pair.

Due to the particularity of raw data (each color dominants one pixel, and local information of area is correlated), by learning and constructing the channel-wise connection between different color channels, we can rearrange channel-wise relationships from the extracted features and reconstruct results of high quality.

3.2. Loss function

Define one short exposure raw image x_{sr} as input and its corresponding long exposure raw image x_{lr}, our network outputs one pseudo long exposure raw image $y_{plr} = f(x_{sr}, \theta)$ and one reconstructed RGB image $y_{rec} = g(x_{sr}, \sigma)$. The Rawpy library[1] offers a process to generate the ground truth RGB image from x_{lr}, which can be represented as a function like $y_{gt} = f(x_{lr}, \omega)$. Our whole pipeline is supervised by Pixel-wise reconstruction loss and Perceptual similarity loss. In the following parts, we discuss these error criterion and their advantages.

3.2.1. Pixel-wise reconstruction loss

The $\mathcal{L}_{rec}$ loss computes error directly on the pixel-level information between the network's output and the ground-truth data. As is found in Zhao, Gallo, Frosio, and Kautz (2017), ℓ_1 can be a simple but effective enough criterion for image reconstruction.

According to our method's architecture, the $\mathcal{L}_{rec}$ consists of supervision between outputs from Short2Long raw restoration subnet and outputs from Color consistent RGB enhancement subnet, *i.e.*the reconstruction error between pseudo raw image data y_{plr} and ground truth raw data x_{lr}, as well as the reconstruction error between final RGB image y_{rec} and ground truth RGB image y_{gt}. The loss function is defined fairly straight-forward as:

$$\mathcal{L}_{rec} = \ell_1(y_{plr}, x_{lr}) + \ell_1(y_{rec}, y_{gt}) \tag{1}$$

3.2.2. Perceptual similarity loss

Single ℓ_1 can result in local optimization solution and affect the visual results, so we introduce a perceptual similarity term to ensure the perceptual change in the structural content of output images to be minimal. The multi-scale structural similarity measure (MS-SSIM) (Wang, Simoncelli, & Bovik, 2003) can evaluate image pairs from perspective of illumination, contrast and structure. We choose MS-SSIM as the perceptual similarity loss.

In order to define the MS-SSIM metric, let us assume a predicted image y_{pred} and its ground truth y_{ori}, we denote μ_{pred}, σ^2_{pred} and $\sigma_{pred,ori}$ as their mean value, variance and covariance, respectively. The structural similarity is computed as:

$$\begin{aligned} &\text{SSIM}(y_{pred}, y_{ori}) \\ &= l(y_{pred}, y_{ori}) * c(y_{pred}, y_{ori}) * s(y_{pred}, y_{ori}) \\ &\approx \frac{(2\mu_{pred}\mu_{ori} + C_1) * (2\sigma_{pred,ori} + C_2)}{(\mu^2_{pred} + \mu^2_{ori} + C_1) * (\sigma^2_{pred} + \sigma^2_{pred} + C_2)}, \end{aligned} \tag{2}$$

where C_1 and C_2 are two relatively small constants. And with $l(y_{pred}, y_{ori})$, $c(y_{pred}, y_{ori})$ and $s(y_{pred}, y_{ori})$, we can compute MS-SSIM as following:

$$\begin{aligned} &\text{MS-SSIM}(y_{pred}, y_{ori}) = [l_m(y_{pred}, y_{ori})]^{\gamma_m} \\ &* \sum_{i=1}^{m} [c_i(y_{pred}, y_{ori})]^{\lambda_i} * [s_i(y_{pred}, y_{ori})]^{\beta_i}, \end{aligned} \tag{3}$$

where γ_m, λ_i and β_i are controllable parameters and m represents the number of scales. For these parameters, we apply the default settings as Wang et al. (2003).

Note that a higher MS-SSIM means a better image. Hence, we need to minimize it as a loss function and extend it to perceptual similarity loss as:

$$\mathcal{L}_{per} = 1 - \frac{1}{N}\sum_{p=1}^{N} \text{MS-SSIM}(y_{pred}, y_{ori}), \tag{4}$$

where y_{pred} and y_{ori} are target and ground truth images.

To evaluate the error from the two branches, we compute it just like the reconstruction loss $\mathcal{L}_{rec}$. By combining two parts of losses, it is formulated as:

$$\begin{aligned} \mathcal{L}_{per} =& (1 - \frac{1}{N}\sum_{p=1}^{N} \text{MS-SSIM}(y_{plr}, x_{lr})) \\ &+ (1 - \frac{1}{N}\sum_{p=1}^{N} \text{MS-SSIM}(y_{rec}, y_{gt})) \end{aligned} \tag{5}$$

Compared with other perceptual loss, our proposed perceptual similarity loss originates from Multi-Scale Structural SImilarity Measure (MS-SSIM) (Wang et al., 2003). It is different from simply calculating the different between extracted features. Here, we get structural similarities from two pairs of sets: (1) GT long exposure raw images and pseudo generated long exposure ones; (2) GT long exposure RGB images and finally generated outputs. The higher MS-SSIMs are, the more similar the GT and generated images are. Considering that perceptual similarity loss is related with MS-SSIM, it can improve the metrics like SSIM.

3.2.3. Total loss

Based on the above sections, we choose a combination of the pixel-wise reconstruction loss and the perceptual similarity loss:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{rec} + (1 - \alpha)\mathcal{L}_{per} \tag{6}$$

where $\alpha \in (0, 1)$ weights the balance of two loss functions. The setting of suitable α is discussed in experiments.

4. Experiments

4.1. Datasets description

We validate methods on three independent low light photography datasets. Two subsets are from Chen et al. (2018).[2] They

[1] https://pypi.org/project/rawpy/

[2] https://cchen156.github.io/SID.html/

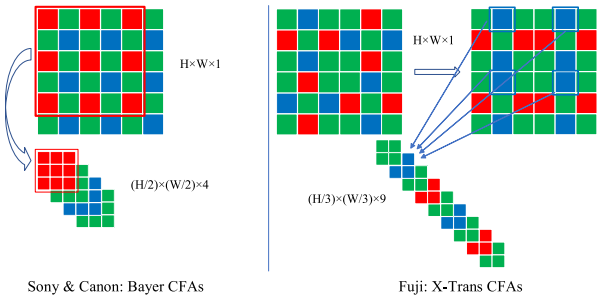


Fig. 4. The illustration of two approaches to packing raw image data.

Table 2
Details of three raw image datasets. From left to right, each column represents: camera type, filter array, exposure time of low exposure raw images, and number of low exposure raw images.

Datasets	Filter array	Exposure time (s)	Number of sets
Sony α7S II	Bayer	1/30	107
		1/25	699
		1/10	1891
Fujifilm X-T2	X-Trans	1/30	630
		1/25	650
		1/10	1117
Canon 6D	Bayer	1/30	122
		1/15	4
		1/10	3
		37	1

are photoed by Sony α7S II and Fujifilm X-T2. The last one[3] is collected from Canon 6D. Sony α7S II is with a Bayer color filter array (CFA) which resolution is 4240 $\times$ 2832, Fujifilm X-T2 is with a X-Trans CFA and 6000 $\times$ 4000 spatial resolution and Canon 6D with a Bayer CFA and 5472 $\times$ 3648 spatial resolution. All three datasets contain 5224 short-exposure raw input images. Note that there are 553 unique long-exposure reference images, indicating that multiple short-exposure input images can correspond to the same ground-truth image. Raw images from Sony α7S II, Fujifilm X-T2 and Canon 6D are all saved in 14 bits, they use different black level values, which are 512, 1024 and 2048, respectively. The details of three raw image datasets are illustrated in Table 2

4.2. Pre-processing details

For the short exposure raw images, they need several pre-processing steps. Considering that only one R/G/B signal locates in each pixel, raw images are packed according to CFAs. Here, we first pack one channel raw image data into four or nine channels according to different CFAs, as is shown in Fig. 4. At the borders of the image sensor, some pixels are never exposed to the light and therefore should be zero (black). However, the values of these pixels are raised due to thermally generated voltage during image acquisition. Hence, we subtract this camera-specific black level bl from the image signal. Besides, n bit raw images need normalization. Hence, for each pixel value x, the processed value x' can be represented as: $x' = \frac{x-bl}{2^n-bl}$. Finally, we consider the pixel-wise value gap of a pair of low and long exposure raw image from perspective of CNNs. Low exposure ones has relatively low numerical value compared with long ones. A reasonable amplification factor should be set to scale the sensor data, which is the ratio between the reference image and the input image. It can be denoted as a simulated value to determine the brightness of output images.

For the corresponding long exposure raw images, we need to generate their RGB images as ground truth. Thanks to the LibRAW[4] library supported by Python, RawPy offers a powerful tool to generate well enough RGB images from well-exposed raw ones. We specific the following parameters in its post-process function: *use_camera_wb=True*, *half_size=False*, *no_auto_bright=True*, *output_bps=16*, and set others as default.

4.3. Training and testing settings

We need to train three networks by those independent datasets. Note that according to Chen et al. (2018) and Chen's explanation in his project page, there are misaligned pairs in Sony dataset, whose ID number are 10034, 10045 and 10172. So we obey the instruction and remove them from the quantitative evaluation stage. For Canon dataset, there are two supplied subsets which are taken under ISO 200 and ISO 12600, separately. We filter a whole set in the same order as Sony and Fuji datasets. We pre-process relevant raw images by settings discussed in Section 4.2, and train networks by computing losses both between the raw data pairs and the RGB image pairs.

All three networks are trained for 4000 epoches. We utilize Adam optimizer with an initial learning rate of 1e−4, which is reduced to 1e−5 after 2000 epoches. All the packed raw images are cropped randomly in 512 $\times$ 512, and randomly rotated or flipped. The batch size is fixed to 1 and we load all the images in memory at first, because the large size of raw images will slow down the dataloader and finally cost more time for training indeed. According to Zhao et al. (2017), we count the numerical scale of Pixel-wise reconstruction loss and Perceptual similarity loss and set $\alpha = 0.16$ to balance two losses.

For testing process, we evaluate on the testing set and images are fed into the network in full size. For each output RGB image, we select its corresponding RGB image which is processed by RawPy to compute metrics.

4.4. Referenced and non-referenced metrics

As is explained in Blau and Michaeli (2018), PSNR and SSIM (Wang, Bovik, Sheikh, & Simoncelli, 2004) are known to correlate less with human visual perception. In certain condition, images with high PSNR and SSIM may have poor perceptual quality. To avoid it, we introduce several recent perceptual quality metrics to give a fairly comparison, *i.e.*PSNR, SSIM and LPIPS (Zhang, Isola, Efros, Shechtman, & Wang, 2018) are three referenced metrics, and NIQE (Mittal, Soundararajan, & Bovik, 2013), LOE (Wang, Zheng, Hu, & Li, 2013) and DeltaE (Sharma, Wu, & Dalal, 2005) are three non-referenced metrics. Hence, six metrics evaluate methods fairly, and the quantitative results can match our visual perception. Note that we add a sign like ↑ or ↓ to indicate whether this metric is positive correlated.

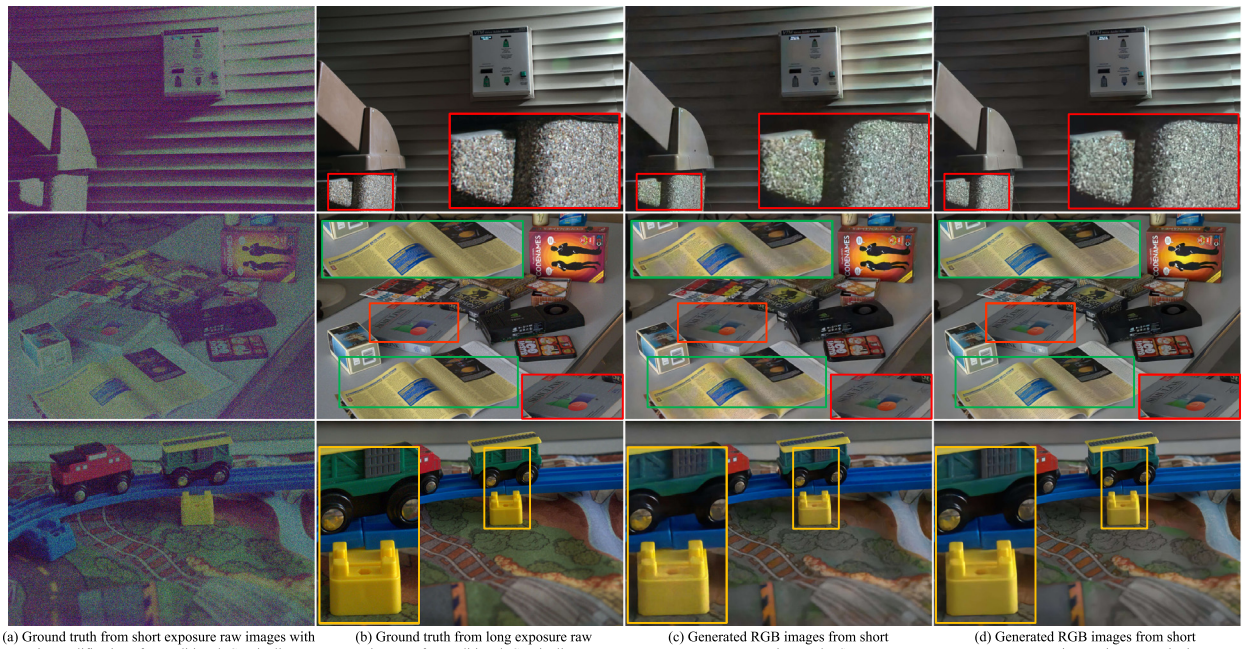
4.5. Qualitative evaluation

Figs. 5–7 show some representative results from three datasets for visual comparison. To give an intuitive impression of different results, we compare the ground truth, the results of SID (Chen et al., 2018) with our method in columns and zoom in details for the fine-grained parts of those images.

Considering that all short exposure raw images are from camera sensors with limited light sensitization, our learning based

[3] https://github.com/jconenna/Canon-6D-Datasets-For-Learning-to-See-in-the-Dark/

[4] https://www.libraw.org/



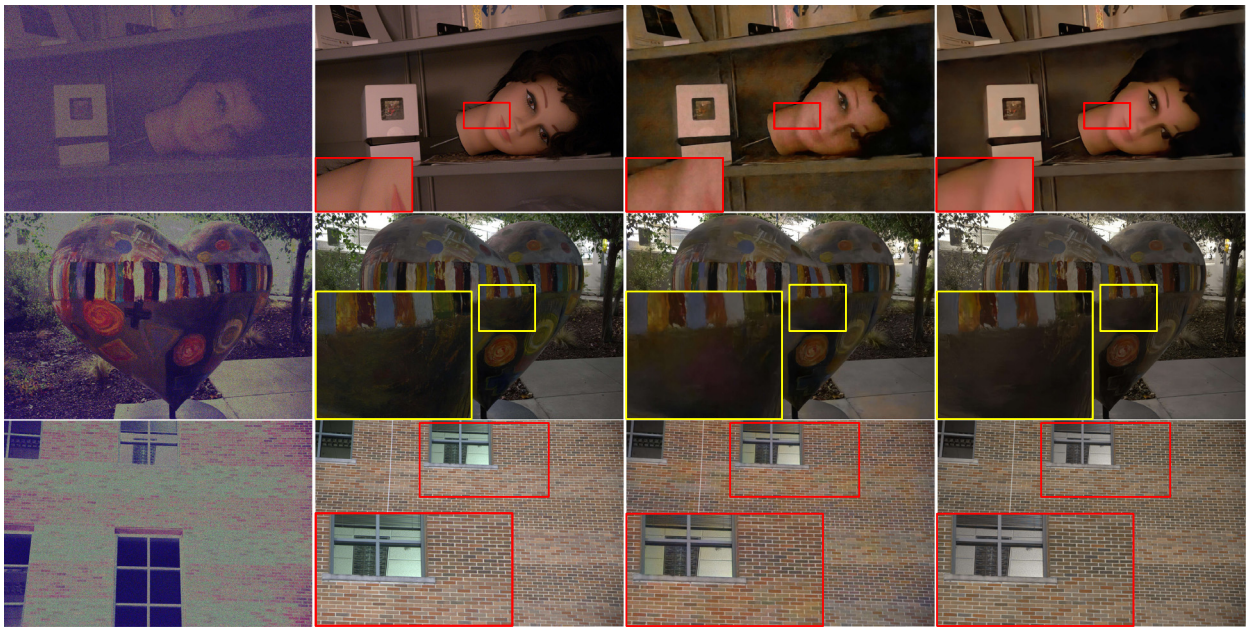
(a) Ground truth from short exposure raw images with ratio amplification after traditional ISP pipeline

(b) Ground truth from long exposure raw images after traditional ISP pipeline

(c) Generated RGB images from short exposure raw images by SID

(d) Generated RGB images from short exposure raw images by our method

Fig. 5. The quality comparisons and zoomed in details of several images from Sony dataset. Each column represents generated RGB images from short exposure raw data by traditional ISP pipelines, Ground Truth from long exposure raw data by traditional ISP pipelines, generated RGB images from short exposure raw data by SID and generated RGB images from short exposure raw data by our method.



(a) Ground truth from short exposure raw images with ratio amplification after traditional ISP pipeline

(b) Ground truth from long exposure raw images after traditional ISP pipeline

(c) Generated RGB images from short exposure raw images by SID

(d) Generated RGB images from short exposure raw images by our method

Fig. 6. The quality comparisons and zoomed in details of several images from Fuji dataset. Each column represents generated RGB images from short exposure raw data by traditional ISP pipelines, Ground Truth from long exposure raw data by traditional ISP pipelines, generated RGB images from short exposure raw data by SID and generated RGB images from short exposure raw data by our method.

method can only try its best to remove those useless noises and learn mappings from short raw data to long data. In other words, our generated pseudo long exposure raw images may lose some details and color information due to its critical data missing. Therefore, we should concentrate on whether those generated RGB images has the color distortion and coarse-grained mosaic. Note that although short exposure raw images can be imaged by traditional ISP pipeline directly, the low level of pixel value results in images all in black. Only with ratio amplification can we get visible results in poor conditions, such as the first column of Fig. 5.

It is clear in Fig. 5 that results from SID suffer a huge color distortion especially on the first and second row. In red boxes of the first row, the stones from SID are colored by green severely compared with our method. In green boxes of the second row, the book page which should be uniformly yellow like the ground truth is converted to several groups of uneven yellow by SID, and our method can recover at most yellow in a more flat way.

(a) Ground truth from short exposure raw images with ratio amplification after traditional ISP pipeline
(b) Ground truth from long exposure raw images after traditional ISP pipeline
(c) Generated RGB images from short exposure raw images by SID
(d) Generated RGB images from short exposure raw images by our method

Fig. 7. The quality comparisons and zoomed in details of several images from Canon dataset. Each column represents generated RGB images from short exposure raw data by traditional ISP pipelines, Ground Truth from long exposure raw data by traditional ISP pipelines, generated RGB images from short exposure raw data by SID and generated RGB images from short exposure raw data by our method.

Table 3
The average of metrics on the Sony dataset.

Method	PSNR ↑	SSIM ↑	NIQE ↓	LOE ↓	DeltaE ↓	LPIPS ↓
CAN (Chen, Xu, & Koltun, 2017)	27.40	0.792	6.61	485.09	8.79	0.637
SID (Chen et al., 2018)	28.88	0.787	7.46	332.77	4.91	0.400
DID (Maharjan et al., 2019)	28.41	0.780	6.95	406.98	8.24	0.465
SGN (Gu, Li, Gool, & Timofte, 2019)	28.91	0.789	7.38	348.57	5.26	0.425
LLPackNet (Lamba, Balaji, & Mitra, 2020)	27.83	0.750	8.42	589.96	9.45	0.614
DCE (Guo et al., 2020)	26.53	0.730	20.95	1426.85	10.79	1.353
REDIIRT (Lamba & Mitra, 2021)	28.66	0.790	6.97	364.69	5.25	0.449
Ours	**29.38**	**0.793**	**6.21**	**313.91**	**4.62**	**0.387**

In red boxes of the second and yellow boxes of the third row, we illustrate the fine-grained details comparison both in text and texture pattern. SID can only recover some structural shapes with severe noises and blurs, but our method generates a clear representation for this kind of detailed areas, and removes noises and blurs to some extend.

In Fig. 6, those are results of Fuji dataset which has a different CFA. Our visual results are still better than other methods. In the first row, our method show obvious contours with few noise. In the second and third row, we can find weird color covered on the box areas generated by SID, while our method has normal color distribution. It reflects the superiority of our method and the expansibility to any CFAs. From Fig. 7, our method shows a better texture preserving ability and can capture nearly correct color.

To conclude, we can find that our results capture natural colors and more details compared with SID and traditional ISP pipeline. The color distortion is relieved by our method. Our results illustrate better characteristics: sharpness, color vividness, good contrast and little noise from low exposure raw images.

4.6. Quantitative evaluation

For a fair evaluation, we collect five recent low exposure enhancement methods, which have no data fusion strategies. Of them, CAN (Chen et al., 2017) is a fully convolutional network for image processing, SID (Chen et al., 2018) is the first attempt for low exposure enhancement on raw data and gets good results, DID (Maharjan et al., 2019) is a residual learning method, SGN (Gu et al., 2019) is designed for RGB images, LLPackNet (Lamba et al., 2020) is a lightweight method with unique pack operations and REDIIRT (Lamba & Mitra, 2021) is a novel method with limited branches of its network. DCE (Guo et al., 2020) is suitable for negligible noise and good color representation.

Due to the lack of experimental statistics for these methods, we reimplement all five methods in PyTorch and perform the same training and testing settings to get their results. By the way, our reimplementation of SID on Sony and Fuji datasets reveals better results than the original report in Chen et al. (2018).

According to 4.4, we adopt the above metrics to evaluate the quantitative quality of reconstructed RGB images. PSNR, SSIM and LPIPS are three referenced metrics, and NIQE, LOE and DeltaE are three non-referenced metrics. For positive correlated metrics, such as PSNR, SSIM, we add a ↑ sign, on the contrast for NIQE, LOE, DeltaE and LPIPS, we represent it with a ↓ sign. All the best scores are in bold for highlight.

It can be seen from Table 3 that our method performs best in every metrics on Sony dataset. For referenced metrics, our method improves 0.47 dB more than the second best method on PSNR and maintains the best SSIM and LPIPS. For other non-referenced metrics, we can observe huge improvements compared with other methods, which correspond with the qualitative results.

Table 4
The average of metrics on the Fuji dataset.

Method	PSNR ↑	SSIM ↑	NIQE ↓	LOE ↓	DeltaE ↓	LPIPS ↓
CAN (Chen et al., 2017)	25.71	0.710	7.41	410.66	8.36	0.719
SID (Chen et al., 2018)	26.61	0.680	**6.31**	387.70	6.43	0.535
DID (Maharjan et al., 2019)	23.81	0.655	6.54	503.73	9.80	0.576
SGN (Gu et al., 2019)	21.58	0.605	8.11	417.41	9.88	0.551
LLPackNet (Lamba et al., 2020)	19.38	0.577	10.60	629.81	11.54	0.969
DCE (Guo et al., 2020)	24.96	0.673	15.33	1358.21	12.03	1.068
REDIIRT (Lamba & Mitra, 2021)	21.40	0.598	8.46	417.79	9.92	0.578
Ours	**27.40**	**0.722**	6.94	**345.19**	**6.00**	**0.505**

Table 5
The average of metrics on the Canon dataset.

Method	PSNR ↑	SSIM ↑	NIQE ↓	LOE ↓	DeltaE ↓	LPIPS ↓
CAN (Chen et al., 2017)	22.89	0.552	7.94	476.11	7.46	0.654
SID (Chen et al., 2018)	28.27	0.655	**7.39**	372.14	5.42	0.519
DID (Maharjan et al., 2019)	27.55	0.621	8.29	379.92	5.88	0.543
SGN (Gu et al., 2019)	21.18	0.538	8.20	576.09	10.57	0.657
LLPackNet (Lamba et al., 2020)	21.52	0.546	13.34	546.21	9.95	0.646
DCE (Guo et al., 2020)	23.84	0.571	9.39	619.92	11.58	0.815
REDIIRT (Lamba & Mitra, 2021)	27.64	0.624	8.88	378.63	5.82	0.529
Ours	**28.41**	**0.662**	7.77	**352.83**	**5.10**	**0.489**

From Table 4, our method is still 0.79 dB ahead of SID on PSNR with good performance on SSIM and LPIPS. However, our method only lead on LOE and DeltaE, and is the third best on NIQE. The main reason of this phenomenon is the domain gap between Sony and Fuji data, because we only model the feature of high quality images from Sony dataset by multivariate Gaussian. The feature representations between Sony and Fuji can be rather different. Also it should be noticed that images from the Fuji come from a totally different CFA unlike Sony and Canon. Hence, CAN, DCE, SGN, LLPackNet and REDIIRT fail on Fuji dataset without specific design of CFA, while our proposed method can be extended to raw data of different structures with a suitable packing strategy.

In Table 5, we can summarize similar conclusions on all metrics. Our method outperforms the others on every metrics except NIQE. The reason of it is similar with results on Fuji dataset. Although Canon 6D has the same structure of CFA like Sony α7S II, the shooting environment and settings of DSLR are different. Hence, the performance of our method on NIQE cannot reach the best due to the domain gap. Here, CAN, DCE, SGN, LLPackNet and REDIIRT also perform bad compared with results on Sony. It mainly results from the small scale of data in Canon dataset.

To summarize, our method show superiority on all three datasets, both on referenced and non-referenced metrics. The quantitative results are consistent with those qualitative demonstrations, *i.e.*sharpness, color vividness, good contrast and little noise from low exposure raw images.

4.7. Ablation study

To evaluate the effectiveness of individual parts in our method, we also perform ablation study on the Sony dataset. As is shown in Table 6, we firstly replace the Color consistent subnet with a simple bi-linear strategy (the first row). The results are close to SID (PSNR 28.88/SSIM 0.787) but still 0.87 dB lower than with the whole pipeline. So it proves the effectiveness of the Color consistent subnet and the efficiency of the Short2Long subnet.

Next, we remove the Short2Long raw restoration subnet to examine the effect of learning pseudo long exposure raw images(the second row). Results reflect that learning from long exposure raw images can remove noises and generate pseudo raw data of high quality. Without this subnet, low quality raw images can affect the training process of the subsequent one — the Color consistent subnet.

Table 6
Ablation study with different subnet combinations on the Sony dataset.

Subnet combinations	PSNR	SSIM
Single Short2Long subnet	28.51	0.787
Single Color consistent subnet	27.40	0.759
DnCNN + Color consistent subnet	27.67	0.773
The whole pipeline	**29.38**	**0.793**

To further examine the denoising ability of the Short2Long raw restoration subnet, we replace it with a classic DnCNN network for comparison(the third row). We can find an increase of 0.27 dB compared with single Color consistent subnet. If we replace it with our Short2Lone subnet (it is actually the full pipeline our method, *i.e.*the last row), we have another 1.71 dB improvement. So we validate the necessity of the Short2Long subnet.

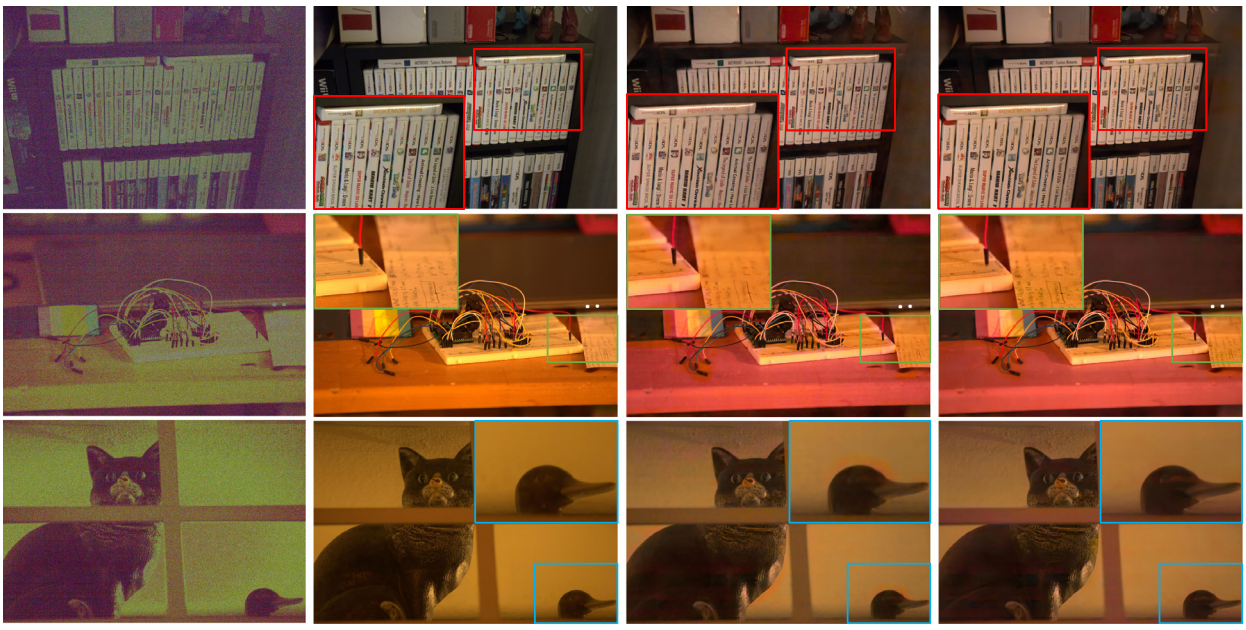
To better verify the effectiveness of each subnet, we illustrate the visualized RGB images in each stage with GT and our whole pipeline in Fig. 10.

4.8. Generalization ability

We verify the generalization capability of the proposed method here. In Fig. 8, the quality comparisons and zoomed in details of several images from Canon dataset are illustrated, we test them with the pre-trained model trained on Sony dataset. It can be observed that our method captures more details and prevents from removing too many textures. Some color distortion areas which can be found from SID are duly handled, so our method generates enhanced images more closer to the original long exposure ones.

What is more, to show the effectiveness of our method in real world applications, we illustrate the quality comparison of the traditional ISP pipelines and ours in Fig. 9. We collect several images shot in low light conditions even nearly dark, both in JPEG (RGB) and DNG (Raw) format. Notice that smartphones will automatically process the received raw data by its traditional ISP pipeline, which is well designed by photography experts, while we only use the raw data as input. So we do nothing to the original JPEG (RGB) images and use corresponding DNG (Raw) images for testing.

Data are from XiaoMi RedMi Note9 and Vivo Z5x. Consider that both of them are the Bayer pattern, to be consistent with



(a) Ground truth from short exposure raw images with ratio amplification after traditional ISP pipeline

(b) Ground truth from long exposure raw images after traditional ISP pipeline

(c) Generated RGB images from short exposure raw images by SID

(d) Generated RGB images from short exposure raw images by our method

Fig. 8. The quality comparisons and zoomed in details of several images from Canon dataset with the pre-trained model from Sony dataset. Each column represents generated RGB images from short exposure raw data by traditional ISP pipelines, Ground Truth from long exposure raw data by traditional ISP pipelines, generated RGB images from short exposure raw data by SID and generated RGB images from short exposure raw data by our method.

(1) Vivo (2) Vivo (3) XiaoMi (4) XiaoMi

Fig. 9. The quality comparisons of the traditional ISP pipelines and ours on XiaoMi and Vivo smartphones. The upper row represents RGB images in JPEG format taken from smartphones. The bottom row represents RGB images generated by our method with raw data input in DNG format.

it, we choose the pre-trained model trained on the Sony dataset. However, due to the differences on sensors, we still need to modify several settings (black level, the order of CFA pattern and bit per sample). From Fig. 9, we can easily distinguish the better ones. Our method can always produce better and satisfying results compared with traditional ISP pipelines for smartphones. It also reflects the strong generalization capability of our method.

5. Discussion and future work

There we discuss the reason why our method can outperform other methods, especially SID. Compared with SID, our method make full use of existing raw data pairs, rather than only considering the simple mapping from short exposure raw images to long exposure RGB images. By learning representations of pseudo long exposure raw images with high quality by the Short2Long raw restoration subnet, our method has already removed noisy points in a certain extent and do no harm the useful information for reconstruction. Then, the subsequent Color consistent RGB enhancement subnet mainly realizes the function of demosaicing, white balance, color correction and gamma correction. It enlightens us on the importance of raw data quality and we should rethink the framework of enhancing low exposure raw data for restoration.

Besides, when removing CCA and the residual structure, the second subnet is degenerated to an architecture like classic DnCNN (Zhang et al., 2017). In DnCNN results, we observe some failure cases of color correction and degradation. This phenomenon reflects that color correction may be harder to learn than the structural representation. Then we focus on our failure cases. Note that there still exist some color distortion on generated RGB images, such as Fig. 5. It is owing to the severe

Fig. 10. The visualization results of our ablation study.

noisy points from original data and the color distortion reflects the pixel-wise differences of noisy and clean area.

Although existing comparable methods on this task are still limited and many difficulties are not solved yet, some related works can be aggregated with our method. Multi exposure fusion (MEF) strategy like EEMEFN (Zhu, Pan, Chen, & Yang, 2020) can fuse frames and improve the restoration quality.

6. Conclusion

In this paper, we propose a joint demosaicing and denoising framework for restoring and enhancing low exposure raw images. It is designed from perspective of learning good representation of raw data and simulating traditional ISP modules, mainly demosaicing and denoising. Without any professional tuning by

experts, we only need several thousand paired raw data. The proposed framework can generate both pseudo raw and RGB outputs, and can be combined with other strategies such as MEF for a better performance. The mid outputs provide high quality pseudo raw data and the final RGB outputs remain the desired attributes: sharpness, color vividness, good contrast and little noise from low exposure raw images. Quantitative and qualitative experimental results prove that our method outperforms against other comparable approaches. The ablation studies validate the effectiveness and efficiency of individual parts of our network. Additional experiments also illustrate the good generalization ability of our method. In the future, we will focus on both accelerating the inference time and improving the model performance, and extend it to video tasks and real world applications.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62122060, 62076188, and the Special Fund of Hubei Luojia Laboratory, PR China under Grant 220100014.

References

Blau, Y., & Michaeli, T. (2018). The perception-distortion tradeoff. In *CVPR* (pp. 6228–6237).
Brooks, T., Mildenhall, B., Xue, T., Chen, J., Sharlet, D., & Barron, J. T. (2019). Unprocessing images for learned raw denoising. In *CVPR* (pp. 11036–11045).
Chatterjee, P., Joshi, N., Kang, S. B., & Matsushita, Y. (2011). Noise suppression in low-light images through joint denoising and demosaicing. In *CVPR* (pp. 321–328).
Chen, C., Chen, Q., Xu, J., & Koltun, V. (2018). Learning to see in the dark. In *CVPR* (pp. 3291–3300).
Chen, Q., Xu, J., & Koltun, V. (2017). Fast image processing with fully-convolutional networks. In *ICCV* (pp. 2516–2525).
Condat, L., & Mosaddegh, S. (2012). Joint demosaicking and denoising by total variationminimization. In *ICIP* (pp. 2781–2784).
Dabov, K., Foi, A., Katkovnik, V., & Egiazarian, K. O. (2007). Image denoising by sparse 3-D transform-domain collaborative filtering. *TIP*, *16*(8), 2080–2095.
Dong, X., Wang, G., Pang, Y., Li, W., Wen, J., Meng, W., et al. (2011). Fast efficient algorithm for enhancement of low lighting video. In *ICME* (pp. 1–6).
Dong, X., Xu, W., Miao, Z., Ma, L., Zhang, C., Yang, J., et al. (2022). Abandoning the bayer-filter to see in the dark. In *CVPR* (pp. 17410–17419).
Ehret, T., Davy, A., Arias, P., & Facciolo, G. (2019). Joint demosaicking and denoising by fine-tuning of bursts of raw images. In *ICCV* (pp. 8867–8876).
Farid, H. (2001). Blind inverse gamma correction. *TIP*, *10*(10), 1428–1433.
Fukushima, T., Kobayashi, Y., Hirasawa, K., Bandoh, T., Ejiri, M., & Kuwahara, H. (1983). An image signal processor. In *IEEE International solid-state circuits conference* (pp. 258–259).
Gharbi, M., Chaurasia, G., Paris, S., & Durand, F. (2016). Deep joint demosaicking and denoising. *ACM Transactions on Graphics*, *35*(6), 191:1–191:12.
Gu, S., Li, Y., Gool, L. V., & Timofte, R. (2019). Self-guided network for fast image denoising. In *ICCV* (pp. 2511–2520).
Guo, C., Li, C., Guo, J., Loy, C. C., Hou, J., Kwong, S., et al. (2020). Zero-reference deep curve estimation for low-light image enhancement. In *CVPR* (pp. 1777–1786).
Guo, X., Li, Y., & Ling, H. (2017). LIME: low-light image enhancement via illumination map estimation. *TIP*, *26*(2), 982–993.
Hirakawa, K., & Parks, T. W. (2006). Joint demosaicing and denoising. *TIP*, *15*(8), 2146–2157.
Ignatov, A., Gool, L. V., & Timofte, R. (2020). Replacing mobile camera ISP with a single deep learning model. In *CVPRW* (pp. 2275–2285).
Jiang, H., Tian, Q., Farrell, J. E., & Wandell, B. A. (2017). Learning the image processing pipeline. *TIP*, *26*(10), 5032–5042.
Jiang, H., & Zheng, Y. (2019). Learning to see moving objects in the dark. In *ICCV* (pp. 7323–7332).
Klatzer, T., Hammernik, K., Knöbelreiter, P., & Pock, T. (2016). Learning joint demosaicing and denoising based on sequential energy minimization. In *ICCP* (pp. 1–11).
Lamba, M., Balaji, A., & Mitra, K. (2020). Towards fast and light-weight restoration of dark images. In *BMVC* (pp. 1–14).
Lamba, M., & Mitra, K. (2021). Restoring extremely dark images in real time. In *CVPR* (pp. 3487–3497).
Lamba, M., & Mitra, K. (2022). Fast and efficient restoration of extremely dark light fields. In *WACV* (pp. 3152–3161).
Liang, Z., Cai, J., Cao, Z., & Zhang, L. (2021). CameraNet: A two-stage framework for effective camera ISP learning. *TIP*, *30*, 2248–2262.
Liang, Z., Guo, S., Gu, H., Zhang, H., & Zhang, L. (2020). A decoupled learning scheme for real-world burst denoising from raw images. *12370*, In *ECCV* (pp. 150–166).
Liu, L., Jia, X., Liu, J., & Tian, Q. (2020). Joint demosaicing and denoising with self guidance. In *CVPR* (pp. 2237–2246).
Liu, Y., Lai, W., Chen, Y., Kao, Y., Yang, M., Chuang, Y., et al. (2020). Single-image HDR reconstruction by learning to reverse the camera pipeline. In *CVPR* (pp. 1648–1657).
Liu, R., Ma, L., Ma, T., Fan, X., & Luo, Z. (2023). Learning with nested scene modeling and cooperative architecture search for low-light vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, *45*(5), 5953–5969.
Liu, Z., Yuan, L., Tang, X., Uyttendaele, M., & Sun, J. (2014). Fast burst images denoising. *ACM Transactions on Graphics*, *33*(6), 232:1–232:9.
Loza, A., Bull, D. R., & Achim, A. (2010). Automatic contrast enhancement of low-light images based on local statistics of wavelet coefficients. In *ICIP* (pp. 3553–3556).
Ma, J., Yan, S., Zhang, L., Wang, G., & Zhang, Q. (2022). ELMformer: Efficient raw image restoration with a locally multiplicative transformer. In *ACM MM* (pp. 5842–5852).
Maharjan, P., Li, L., Li, Z., Xu, N., Ma, C., & Li, Y. (2019). Improving extreme low-light image denoising via residual learning. In *ICME* (pp. 916–921).
Malm, H., Oskarsson, M., Warrant, E., Clarberg, P., Hasselgren, J., & Lejdfors, C. (2007). Adaptive enhancement and noise reduction in very low light-level video. In *ICCV* (pp. 1–8).
Mittal, A., Soundararajan, R., & Bovik, A. C. (2013). Making a "Completely blind" image quality analyzer. *SPL*, *20*(3), 209–212.
Morawski, I., Chen, Y., Lin, Y., Dangi, S., He, K., & Hsu, W. H. (2022). Genisp: Neural ISP for low-light machine cognition. In *CVPR Workshops* (pp. 629–638).
Park, S., Yu, S., Moon, B., Ko, S., & Paik, J. (2017). Low-light image enhancement using variational optimization-based retinex model. *TCE*, *63*(2), 178–184.
Ren, J., Zhang, Z., Hong, R., Xu, M., Zhang, H., Zhao, M., et al. (2022). Robust low-rank convolution network for image denoising. In *ACM MM* (pp. 6211–6219).
Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *MICCAI* (pp. 234–241).
Schwartz, E., Giryes, R., & Bronstein, A. M. (2019). Deepisp: Toward learning an end-to-end image processing pipeline. *TIP*, *28*(2), 912–923.
Sharma, G., Wu, W., & Dalal, E. N. (2005). The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application: Endorsed By Inter-Society Color Council, the Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, the Swedish Colour Centre Foundation, Colour Society of Australia, Centre FranÇAis de la Couleur*, *30*(1), 21–30.
Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., et al. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR* (pp. 1874–1883).
Stark, J. A. (2000). Adaptive image contrast enhancement using generalizations of histogram equalization. *TIP*, *9*(5), 889–896.
Tan, H., Zeng, X., Lai, S., Liu, Y., & Zhang, M. (2017). Joint demosaicing and denoising of noisy bayer images with ADMM. In *ICIP* (pp. 2951–2955).
Tian, C., Xu, Y., Li, Z., Zuo, W., Fei, L., & Liu, H. (2020). Attention-guided CNN for image denoising. *Neural Networks*, *124*, 117–129.
Tian, C., Yuan, Y., Zhang, S., Lin, C., Zuo, W., & Zhang, D. (2022). Image super-resolution with an enhanced group convolutional neural network. *Neural Networks*, *153*, 373–385.
Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *TIP*, *13*(4), 600–612.
Wang, Y., Huang, H., Xu, Q., Liu, J., Liu, Y., & Wang, J. (2020). Practical deep raw image denoising on mobile devices. *12351*, In *ECCV* (pp. 1–16).
Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *Asilomar conference on signals, systems computers, vol. 2* (pp. 1398–1402 Vol.2).
Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In *CVPR* (pp. 11531–11539).

Wang, S., Zheng, J., Hu, H., & Li, B. (2013). Naturalness preserved enhancement algorithm for non-uniform illumination images. *TIP*, *22*(9), 3538–3548.
Wei, K., Fu, Y., Yang, J., & Huang, H. (2020). A physics-based noise formation model for extreme low-light raw denoising. In *CVPR* (pp. 2755–2764).
Xing, W., & Egiazarian, K. O. (2021). End-to-end learning for joint image demosaicing, denoising and super-resolution. In *CVPR* (pp. 3507–3516).
Xu, K., Yang, X., Yin, B., & Lau, R. W. H. (2020). Learning to restore low-light images via decomposition-and-enhancement. In *CVPR* (pp. 2278–2287).
Xu, S., Zhang, C., & Zhang, J. (2020). Bayesian deep matrix factorization network for multiple images denoising. *Neural Networks*, *123*, 420–428.
Yang, W., Wang, S., Fang, Y., Wang, Y., & Liu, J. (2020). From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *CVPR* (pp. 3060–3069).
Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR* (pp. 586–595).
Zhang, L., Song, L., Du, B., & Zhang, Y. (2021). Nonlocal low-rank tensor completion for visual data. *TCYB*, *51*(2), 673–685.
Zhang, Z., Zheng, H., Hong, R., Xu, M., Yan, S., & Wang, M. (2022). Deep color consistent network for low-light image enhancement. In *CVPR* (pp. 1889–1898).
Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *TIP*, *26*(7), 3142–3155.
Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2017). Loss functions for image restoration with neural networks. *TCI*, *3*(1), 47–57.
Zhu, M., Pan, P., Chen, W., & Yang, Y. (2020). EEMEFN: low-light image enhancement via edge-enhanced multi-exposure fusion network. In *AAAI* (pp. 13106–13113).