Abandoning the Bayer-Filter to See in the Dark

Xingbo Dong[1,3*†] Wanyan Xu[1,2*†] Zhihui Miao[1,2†] Lan Ma[1]
Chao Zhang[1] Jiewen Yang[1] Zhe Jin[4] Andrew Beng Jin Teoh[3] Jiajun Shen[1]
[1]TCL AI Lab [2]Fuzhou University [3]Yonsei University [4]Anhui University
{xingbo.dong,bjteoh}@yonsei.ac.kr,{208527051,208527090}@fzu.edu.cn,{sjj,rubyma}@tcl.com

Abstract

Low-light image enhancement, a pervasive but challenging problem, plays a central role in enhancing the visibility of an image captured in a poor illumination environment. Due to the fact that not all photons can pass the Bayer-Filter on the sensor of the color camera, in this work, we first present a De-Bayer-Filter simulator based on deep neural networks to generate a monochrome raw image from the colored raw image. Next, a fully convolutional network is proposed to achieve the low-light image enhancement by fusing colored raw data with synthesized monochrome data. Channel-wise attention is also introduced to the fusion process to establish a complementary interaction between features from colored and monochrome raw images. To train the convolutional networks, we propose a dataset with monochrome and color raw pairs named Mono-Colored Raw paired dataset (MCR) collected by using a monochrome camera without Bayer-Filter and a color camera with Bayer-Filter. The proposed pipeline takes advantages of the fusion of the virtual monochrome and the color raw images, and our extensive experiments indicate that significant improvement can be achieved by leveraging raw sensor data and data-driven learning. The project is available at https://github.com/TCL-AILab/Abandon_Bayer-Filter_See_in_the_Dark

1. Introduction

For a digitalized image, the quality of the image could be severely degraded due to the color distortions and noise under poor illumination conditions such as indoors, at night, or under improper camera exposure parameters.

Long exposure time and high ISO (sensitivity to light) are often leveraged in low-light environments to preserve visual quality. However, overwhelming exposure leads to motion blur and unbalanced overexposing, and high ISO amplifies the noise. Though the camera's flash provides exposure compensation for the insufficient light, it is not suitable for long-distance shots, and also introduces color distortions and artifacts. On the other hand, various algorithms have been reported to enhance the low-light image. Recently, deep neural network models have been utilized to solve the low-light image restoration problem, such as DeepISP [22] and Seeing In the Dark (SID) [3].

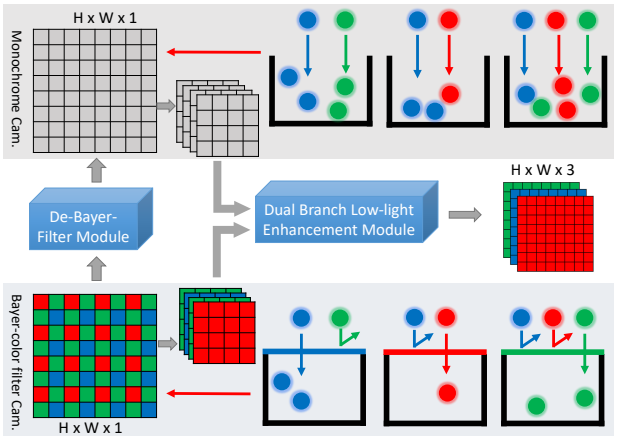


Figure 1. Overview of the proposed pipeline. We propose to generate monochrome raw data by a learned De-Bayer-Filter module. Then, a dual branch neural network is designed to bridge monochrome and colored raw to achieve the low-light image enhancement task.

However, those algorithms are restricted in the image processing pipeline, as the photons capture rate and quantum efficiency are usually overlooked. In general, high photons capture rate can improve the image's visual quality significantly. One of the typical examples is the RYYB-based color filter, which can capture 40% more photons than the Bayer-RGGB-based color filter[1]. Hence, the RYYB-based color filter can achieve better performance naturally.

Bayer filter removal is another plausible way to improve the photons capture rate. The Bayer filter is an array of

*These authors contributed equally to this work.
†Work done while interning at TCL AI Lab.

[1]Bayer filter, Bayer-array, Bayer-array filter are used interchangeably.

many tiny color filters that cover the image sensor to render color information (see Fig. 1). By removing the Bayer filter and sacrificing the color information, the image sensor can capture more photons, which contributes to clearer visibility under poor illumination conditions compared to a camera with a Bayer filter (see Fig. 2 (a)). On the other hand, dual-cameras are one of the trends of today's smart devices such as smartphones. One type of dual-camera set is the combination of monochrome sensor and colored sensor[2]. The monochrome sensor is usually identical to the colored sensor but without a Bayer array filter. Such a dual-camera setting can achieve better imaging quality in a low-light environment due to more photons received by the sensor. However, an additional cost is needed for the extra camera equipped. Therefore, for most mobile phones that are only equipped with color cameras, preserving the same low-light image quality produced by dual-camera set while only using a single color camera is a challenging task.

Motivated by the above discussion, we proposed a fully end-to-end convolutional neural model that consists of two modules (as illustrated in Fig. 1): a De-Bayer-Filter (DBF) module and a Dual Branch Low-light Enhancement module (DBLE). The DBF module learns to restore the monochrome raw image from the color camera raw data without requiring a monochrome camera. DBLE is designed to fuse colored raw with synthesized monochrome raw data and generate enhanced RGB images.

In addition, we propose a dataset to train our end-to-end framework. To the best of our knowledge, no existing dataset contains monochrome and colored raw image pairs captured by an identical type of sensors. To establish such a dataset, one camera with a Bayer filter is used to capture color-patterned raw images. Another camera without a Bayer-filter but equipped with the same type of sensor is utilized to capture monochrome raw images (see Fig. 2 (b)). The dataset is collected under various scenes, and each colored raw image has a corresponding monochrome raw image captured with identical exposure settings.

Our contributions can be summarised as:

1. A De-Bayer-Filter model is proposed to simulate a virtual monochrome camera and synthesize monochrome raw image data from the colored raw input.The DBF module aims at predicting the monochrome raw images, which resembles a monochrome sensor capability. To the best of our knowledge, we are the first to explore removing the Bayer-filter using a deep learning-based model.

2. We design a Dual Branch Low-light Enhancement model that is used to fuse the colored raw with the synthesized monochrome raw to produce the final monitor-ready RGB images. To bridge the domain gap between colored raw and monochrome raw, a channel-wise attention layer is adopted to build an interaction between both domains for better restoration performance. The experiment results indicate that state-of-the-art performance can be achieved.

3. We propose the **MCR**, a dataset of colored raw and monochrome raw image pairs, captured with the same exposure setting. It is publicly opened as a research material to facilitate community utilization and will be released after publication.

[2]For example, Huawei P9, Moto Z2 Force

2. Related Work

To achieve the low-light image enhancement task, tremendous methods have been attempted. These methods can be categorized as histogram equalization (HE) methods [1, 15, 29], Retinex methods [5, 26, 28, 33], defogging model methods [4], statistical methods [16, 17, 23], and machine learning methods [7, 11, 30, 34]. Recently, several works on raw image data have been proposed [3, 9, 22]. Our work also falls into this category; we will mainly discuss the existing methods of raw-based approaches in this section.

Deep neural networks have emerged as an approach to achieve the digital camera's image signal processing tasks. In 2018, a fully convolutional model, namely DeepISP, was proposed in [22] to learn mapping from the raw low-light mosaiced image to the final RGB image with high visual quality. To simulate the digital camera's image signal processing (ISP) pipeline, DeepISP first extracts low-level features and performs local modifications, then extracts higher-level features and performs a global correction. L1 norm and the multi-scale structural similarity index (MS-SSIM) loss in the Lab domain are utilized for training the DeepISP to simulate the ISP pipeline. When DeepISP is only used for low-level imaging tasks such as denoising and demosaicing, L2 loss will be utilized. Hence, both low-level tasks and higher-level tasks such as demosaicing, denoising, and color correction can be achieved by DeepISP. The results in [22] suggest superior performance compared with manufacturer ISP.

Another parallel work similar to DeepISP, namely seeing in the dark (SID), was proposed in [3]. In SID, a U-net [21] network is utilized to operate directly on raw sensor data and output human visual ready RGB images. A dataset of raw short-exposure low-light images with corresponding long-exposure reference images was established to train the model. Compared with the traditional image processing pipeline, significant improvement can be made as the results in [3] indicate. Later, an improved version of SID was proposed in [27]. Using a similar U-net network as the backbone, the authors introduced wavelet transform to conduct down-sampling and up-sampling operations. Perceptual loss [10] is used in [27] to train the network to better

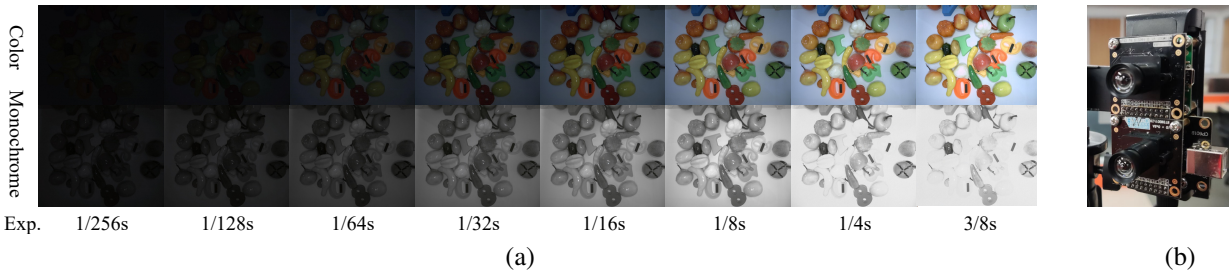


Figure 2. (a) Images captured by color and monochrome cameras under different exposure time.; (b) Monochrome and color cameras used in our work for data collection.

restore details in the image. In DID [18], the authors proposed replacing the U-net in SID with residual learning to better preserve the information from image features. Similar raw-based approaches have also been applied to videos, such as [2, 9].

In addition to the raw-based approach, frequency-based decomposition has also been explored on the low-light image enhancement task. In [31], the authors proposed a pipeline, namely LDC, to achieve the low-light image enhancement task based on a frequency-based decomposition and enhancement model. The model first filters out high-frequency features and learns to restore the remaining low-frequency features based on an amplification operation. Subsequently, high-frequency details are restored. The results from [31] indicate that state-of-the-art performance can be achieved by LDC.

Various research has also been done to improve the efficiency of low-light image enhancement in raw domain. To achieve a computationally fast low-light enhancement system, the authors in [14] proposed a lightweight architecture (RED) for extreme low-light image restoration. Besides, the authors also proposed an amplifier module to estimate the amplification factor based on the input raw image. In [6], a self-guided neural network (SGN) was proposed to achieve a balance between denoising performance and the computational cost. It aims at guiding the image restoration process at finer scales by utilizing the large-scale contextual information from shuffled multi-resolution inputs.

Methods discussed above generally learn to map raw data captured by the camera to the human-visual-ready image. As raw data provides full information, the reviewed approach achieves state-of-the-art performance. However, the performance of those methods is upper bounded by the information contained in the raw data. While in our work, we consider to introduce extra information beyond the raw-RGB data.

3. The Method

Motivated by the above discussion and inspired by the monochrome camera's high light sensitivity, we propose a novel pipeline to further push the raw-based approaches forward. Specifically, our pipeline takes a raw image captured by a color camera with a Bayer-Filter as input. The De-Bayer-Filter module in our pipeline will first generate a monochrome image; a dual branch low-light enhancement module then fuses the monochrome raw data and color raw data to produce the final enhanced RGB image. Both modules work on raw images, as raw images are linearly dependent on the number of photons received, which contains additional information compared to RGB images such as the noise distribution [2, 20]. Details of each module will be discussed subsequently. A detailed architecture diagram of our framework is shown in Fig. 3(a) (more details are discussed in the supplementary). Furthermore, Fig. 3(b-f) and Fig. 3(g-k) visualize the output of each step of our model on our dataset and the SID dataset in [3], respectively.

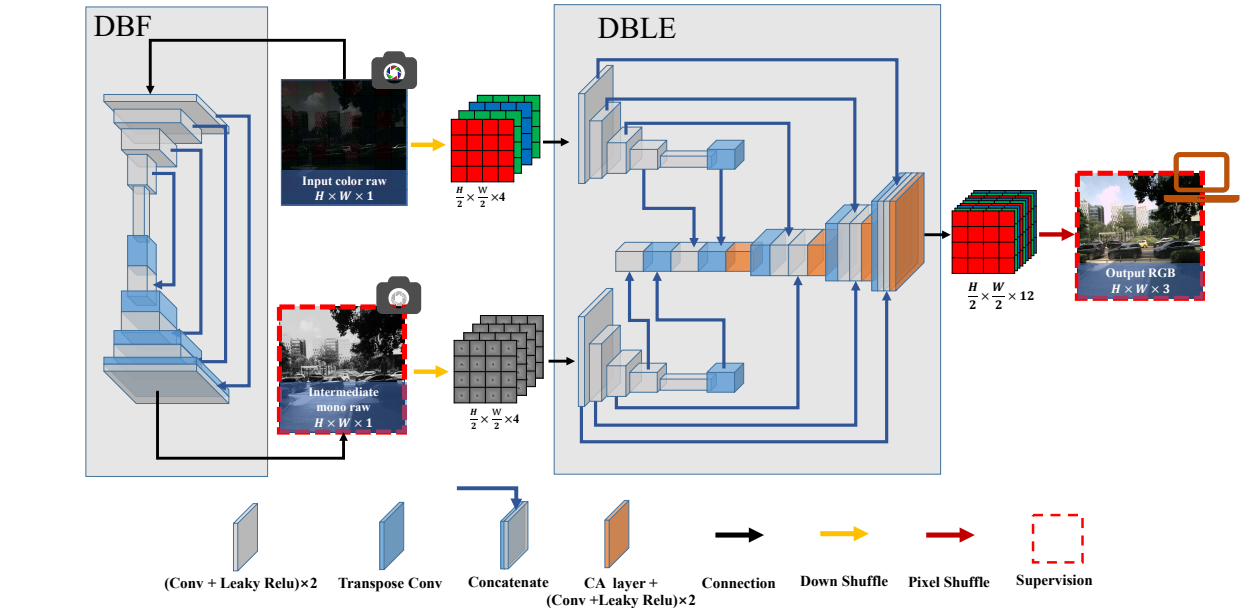
3.1. De-Bayer-Filter Module

Millions of tiny light cavities are designed to collect photons and activate electrical signals on the camera sensor. However, using those light cavities alone can only produce gray images. A Bayer color filter is therefore designed to cover the light cavities and collect color information to produce color images. More specifically, a standard Bayer unit is a 2×2 pixel block with two green, one red and one blue color filters, and filters of a certain color will only allow photons with the corresponding wavelength to pass through.

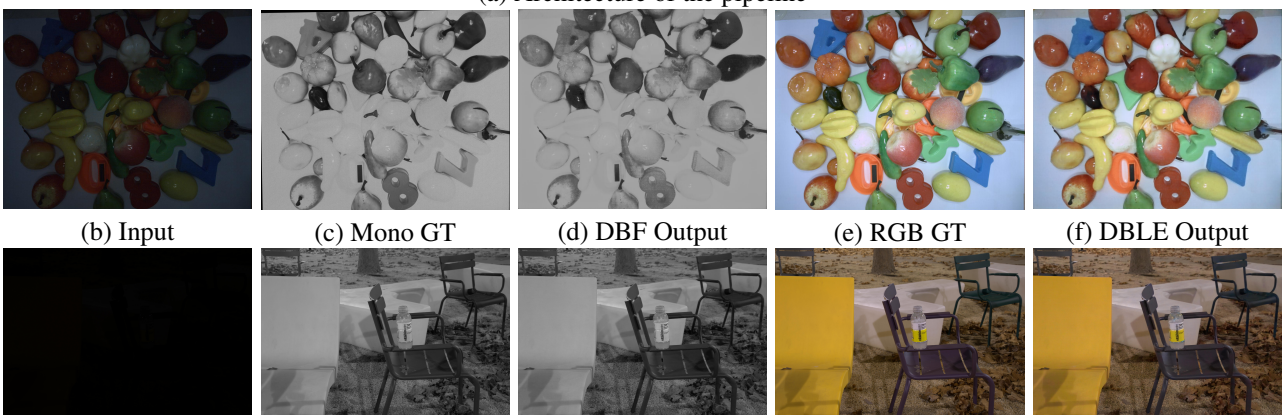
Simulating the camera imaging process using neural networks has been demonstrated feasible in several works [3,20,22]. Inspired by those works, we consider the removal of the Bayer array filter virtually by modeling the relationship between input and output photons for each color filter. Specifically, a De-Bayer-Filter (DBF) module is designed in this work to restore the monochrome raw images $A_{mono} \in \mathbb{R}^{H \times W}$ from the input colored raw $A_{color} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 4}$:

$$A_{Mono} = f_M(A_{Color}) \tag{1}$$

where $f_M(\cdot)$ is a U-net-based fully convolutional network (see Fig. 3). L1 distance between the ground-truth monochrome image A^{GT}_{Mono} and predicted image A_{Mono}



(a) Architecture of the pipeline



(b) Input (c) Mono GT (d) DBF Output (e) RGB GT (f) DBLE Output

(g) Input (h) Synthetic Mono GT (i) DBF Output (j) RGB GT (k) DBLE Output

Figure 3. (a) is the architecture of the pipeline. DBF module is designed to produce a monochrome image from the input raw image. DBLE module is proposed to fuse color and monochrome raw images to enhance the low-light input image. Each box denotes a multi-channel feature map produced by each layer. (b)-(f) are the images of our pipeline trained on our dataset. (g)-(k) are the images of our pipeline trained on SID [3] dataset; we convert RGB ground truth (GT) in SID dataset to gray image to replace the monochrome GT in our dataset.

is used as a loss to encourage the DBF to learn to restore monochrome images with more details from low-light raw images. We hypothesize that the generated monochrome raw image can enhance the low-light image by introducing more information into the subsequent module.

3.2. Dual Branch Low-Light Image Enhancement Module

There are many differences between the colored raw image and monochrome image: 1) colored raw images have mosaic patterns; 2) the colored raw images consist of four channels with a resolution of $\frac{H}{2} \times \frac{W}{2}$, while their counterparts consist of one channel with $H \times W$ resolution; 3) no color information is included in the monochrome images; 4) better illuminating information is preserved on monochrome images as the monochrome camera sensor can better capture the light.

Based on the above observations, we propose a dual branch low-light image enhancement (DBLE) module (see Fig. 3), which treats the DBF generated monochrome raw image and colored raw image separately in the down-sampling process. Meanwhile, different level feature maps of the two down-sampling branches are fused based on concatenation and followed by channel-wise attention (CA) layer [8] in the up-sampling branch to synthesize the human-visual ready RGB image $I_{rgb} \in \mathbb{R}^{H \times W \times 3}$. The

DBLE module is defined as:

$$I_{RGB} = f_C(A_{Color}; A_{Mono}), \tag{2}$$

where f_C is a specifically designed fully convolutional network, which is shown in Fig. 3 (a). L1 distance between the ground truth RGB image I^{GT}_{RGB} and predicted image I_{RGB} is used as the loss to encourage the DBLE to learn to restore visual-ready RGB output from low-light raw images.

As the conventional U-net network treats features from each channel equally, directly concatenating the feature map from the monochrome raw branch and colored raw branch may lead to contradiction due to the domain gap. The usage of strided convolution and transposed convolution layers will also lead to spatial information loss. Motivated by [32], after the concatenation operation, a CA layer [8] is adopted to achieve a channel-wise attention recalibration in DBLE to bridge the gap between monochrome and color images. The CA layer can explicitly model the interaction of colored raw and monochrome raw modalities to exploit the complementariness and reduce contradiction from both domains.

It has been reported that upsampling layers (transposed convolutional layers) used in U-net causes images to be distorted by checkerboard artifacts [13, 19, 24, 25]. We also found such checkerboard artifacts in our settings on U-net, especially for images with white backgrounds. In our work, the CA layer also serves a role in avoiding checkerboard artifacts. As downscale and upscale operations are included in the CA layer, the CA layer is similar to the resize-convolution operation which discourages high-frequency artifacts in a weight-tying manner [19].

3.3. Dataset Design

Mono-Colored Raw Paired (MCR) Dataset. To the best of our knowledge, no existing dataset contains monochrome and Bayer raw image pairs captured by the same type of sensors. To establish the dataset, we capture image pairs of the same scenes with two cameras, denoted as Cam-Color and Cam-Mono[3]. Both cameras have the same 1/2-inch CMOS sensor and output a 1,280H x 1,024V imaging pixel array. However, only Cam-Color is equipped with a Bayer color filter. Cam-Color is used to capture colored raw images in our work, and Cam-Mono captures monochrome raw images.

We collect the data in both indoor and outdoor conditions. The illuminance at the indoor scenes is between 50 lux and 2,000 lux under regular lights. The outdoor images were captured during daytime and night, under sun lighting or street lighting, with an illuminance between 900 lux and 14,000 lux. The captured scenes includes toys, books, stationery objects, street views, and parks.

[3]Part Number: MT9M001C12STC/MT9M001C12STM

Table 1. Summary of the dataset

Scenes	Exposure time (s)	Data Pairs	Fixed Settings
Indoor fixed position	1/256, 1/128, 1/64, 1/32, 1/16, 1/8, 1/4, 3/8	2744 pairs	Format: .raw, resolution: 1280*1024
Indoor sliding platform	1/256, 1/128, 1/64, 1/32, 1/16, 1/8, 1/4, 3/8	800 pairs	
Outdoor sliding platform	1/4096, 1/2048, 1/1024, 1/512, 1/256, 1/128, 1/64, 1/32	440 pairs	

The cameras are mounted on the sliding platform on sturdy tripods or a fixed platform on a sturdy table. When mounted on the sliding platform, the camera is adjusted to the same position by sliding the platform to minimize the position displacement among images captured by two cameras in the same scene. When mounted on the fixed platform, the camera is attached to the same position as the platform to minimize the position displacement. Camera gain is set with the camera default value. Focal lengths are adjusted to maximize the quality of the images under long exposure. The exposure time is adjusted according to the specific scene environment.

Position displacement is unavoidable in the capture process. Hence, it is necessary to align the images captured from two cameras. The best exposure colored raw and monochrome raw is selected to align the images captured by two cameras in the same scenes. Then, homography feature matching is utilized to extract key points from the selected image pair, and a brute force matcher is utilized to find the matched key points. The extracted locations of good matches are filtered based on an empirical thresholding method. A homography matrix can be decided based on the filtered location of good matches. Finally, the homography transformation is applied to other images captured from the same scene. The statistic information of the dataset is summarized in Table 1. Fig. 2(a) demonstrates a series of monochrome-colored raw paired images from the dataset.

Artificial Mono-Colored Raw SID Dataset. The original SID dataset collected in [3] contains 5,094 raw short-exposure images taken from the indoor and outdoor environments, while each short-exposure image has a corresponding long-exposure reference image. The short exposure time is usually between 1/30 second and 1/10 second, and the exposure time of the corresponding long-exposure image is 10 to 30 seconds.

However, monochrome images are not available in the original SID dataset. To address this, we built an artificial Mono-colored raw dataset based on SID [3] dataset in this work. More specifically, we first convert the long-exposure raw images in the original SID dataset to RGB images, and these RGB images are further converted to grayscale by forming a weighted sum of the R, G, and B channels, as shown in Fig. 3(h). Such conversion can eliminate the hue and saturation information while retaining the luminance information.

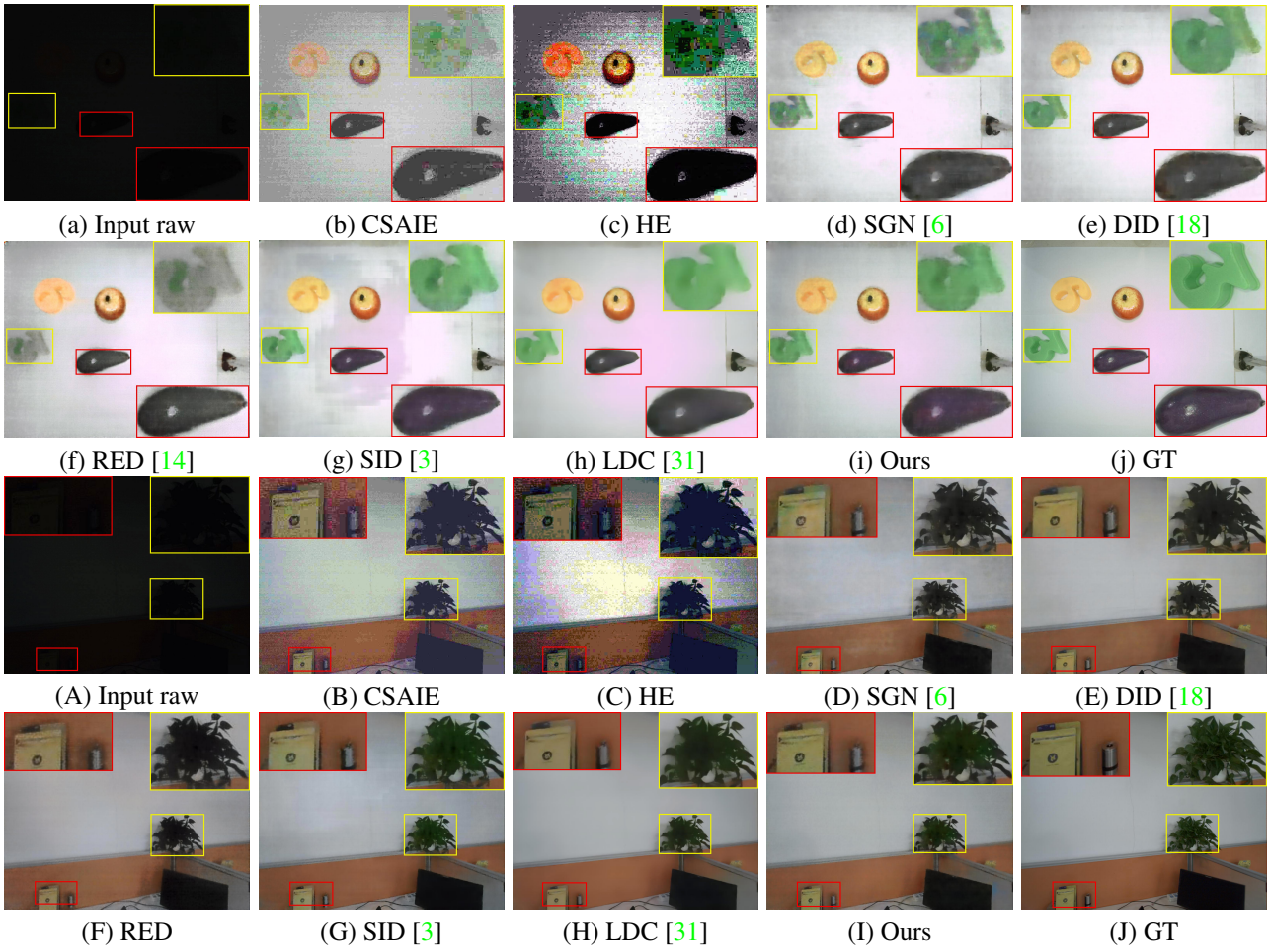


Figure 4. Visual results of state-of-the-art methods and ours on low-light images RAW in our dataset. The larger boxes show the zoom-in version of the regions in the smaller boxes of the same color. The 'CSAIE' means 'Commercial Software Automatic Image Enhancement'.

3.4. Training

By default, we pre-process the input images similarly to [3] where images' pixel values are amplified with predefined ratios followed by a pack raw operation. We incorporate the CA layer [8] to bridge the domain gap between features from monochrome and colored raw images. The whole system is trained jointly with L1 loss to directly output the corresponding long-exposure monochrome and sRGB images.

The dataset is split into train and test sets without overlapping by the ratio of 9:1. The input patches are randomly cropped from the original images with 512 × 512. In the case of raw image input, the RGGB pixel position is carefully preserved in the cropping process. We implement our model with Pytorch 1.7 on the RTX 3090 GPU platform, and we train the networks from scratch using the Adam [12] optimizer. The learning rate was set to 10^{-4} and 10^{-5} after converging, and the weight-decay was set to 0.

4. Experiments and Results

In this section, we present a comprehensive performance evaluation of the proposed low-light image enhancement system. To measure the performance, we evaluate the system performance in terms of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). For PSNR and SSIM, a higher value means a better similarity between output image and ground truth.

4.1. Comparison with State-of-the-Art Methods

Qualitative Comparison. We first visually compare the results of the proposed method with other state-of-the-art deep learning-based image enhancement methods, including SID [3], DID [18], SGN [6], LDC [31], and RED [14]. In addition, the traditional histogram equalization (HE) approach and a Commercial Software Automatic Image Enhancement (CSAIE) method are also included in the com-

Table 2. Comparison with SOTA.

	MCR Dataset		SID Dataset	
	PSNR (dB)	SSIM	PSNR (dB)	SSIM
RED [14] *(21,CVPR)*	25.74	0.851	28.66	0.790
SGN [6] *(19,ICCV)*	26.29	0.882	28.91	0.789
DID [18] *(19,ICME)*	26.16	0.888	28.41	0.780
SID [3] *(18,CVPR)*	29.00	0.906	28.88	0.787
LDC [31] *(20,CVPR)*	29.36	0.904	29.56	**0.799**
Ours	**31.69**	**0.908**	**29.65**	0.797

parison. Fig. 4 shows the results of different methods on two low-light images (see more results in supplementary).

As indicated by Fig. 4, our method can achieve better enhancement and denoising visual performance. Specifically, checkerboard artifacts are usually found on SID for images with white background. This is because of the usage of upsampling layers in the model. Foggy artifacts are usually observed on SGN; color distortions also are found on SGN, DID, and RED, as are shown in Fig. 4 (A-J), where the green plant enclosed by the yellow box becomes black after restoring by SGN, DID, and RED. Compared to LDC, our methods can preserve more details as over-smoothing is usually found on LDC. Note that over-smoothing may be more visual appealing, but details will be lost, for example, the wall crack becomes invisible on LDC as shown in Fig. 4 (H-I). In a nutshell, Fig. 4 demonstrates the satisfying visual performance achieved by our method, with fewer artifacts but more convincing restoration.

Quantitative Comparison. A quantitative comparison against the state-of-the-art enhancement methods has also been performed. For a fair comparison, SID [3], DID [18], SGN [6], LDC [31], and RED [14] were trained on the MCR dataset.

As Table 2 shows, our proposed method outperforms its counterparts by a large margin. Specifically, our method can achieve a PSNR of 31.69dB on MCR dataset, which is 7.9% higher than the second-best method, i.e., the LDC [31]. Our method can also achieve an SSIM of 0.908, which is the highest among all compared methods.

Compared to other methods, we incorporate the extra monochrome information into the processing pipeline, hence state-of-the-art performance can be achieved. As shown in the first two data rows in Table 2, both RED [14] and SGN [6] can only achieve a PSNR of around 26dB. Both RED and SGN aim at reducing the computational cost and improving efficiency. Hence it is reasonable to observe the performance degradation. The result on DID [18] from Table 2 suggests that replacing U-net with residual learning cannot achieve superior performance on our dataset.

On the MCR dataset, SID [3] achieves a PSNR of only 29.00dB. The checkerboard artifact may be the reason. From Table 2, we observe that LDC [31] achieves the second-best performance. This is because they are based on a frequency-based decomposition and enhancement model, which can better restore the noisy image and avoid noise amplification. We also train our model on the modified SID dataset to further validate our method for a fair comparison. The performance results are shown in the SID column in Table 2. As the results suggest, our method also outperforms all its counterparts. Specifically, our method can achieve a PSNR of 29.65dB, which is around 0.1dB higher than LDC, while the SSIM can achieve similar performance.

Other methods including SID, DID, SGN, and RED can only achieve a PSNR around 28dB. In summary, the results show that our model is more effective in enhancing low-light images with noise. The performance of most existing methods is upper bounded by the information contained in the raw data. In our proposed pipeline, we further extend the upper bound by considering the monochrome domain. Hence, better performance can be achieved.

4.2. Ablation Studies

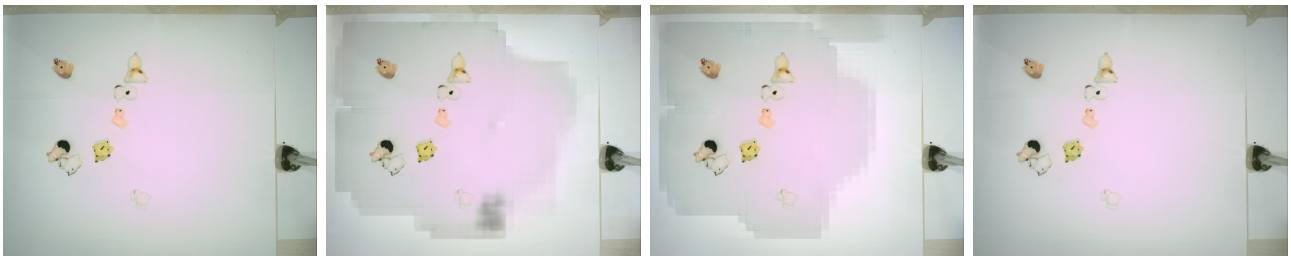
In this subsection, we provide several ablation studies for the proposed system to better demonstrate the effectiveness of each module of our system.

Checkerboard artifacts are found in our preliminary exploration stage, especially for images with white backgrounds. To eliminate checkerboard artifacts, we incorporate the CA layer [8] in the DBLE module. In this ablation study, we first remove the CA layer in the DBLE module to demonstrate the checkerboard artifacts' elimination and performance upgrading. Besides, we also train an original SID [3] network on our dataset to show the visual effect of the checkerboard artifacts of U-net. The restored images from SID, DBLE without CA layer, and DBLE with CA layer are shown in Fig. 5. It is observed that checkerboard artifacts can be perfectly avoided by introducing the CA layer. Besides, as per the quantitative results shown in Table 3, CA layer can boost the image enhancement performance as the PSNR increases to 31.69dB compared with its counterpart of 29.23dB.

We also train the model to learn the ratio directly instead of amplifying image pixel values with predefined ratios. Hence, we train a model without amplifying the input raw images with the predefined ratio. As a result, as shown in Table 3, such a model can still achieve comparable performance, with only a slight decrease in PSNR and SSIM.

As suggested by [3], we change the packraw-based input into original one-channel raw images. As shown in the row of baseline without packraw in Table 3, PSNR and SSIM degradation is observed. We argue that the packing of raw can assist the model to better process the color information.

The change of loss function from L1 to L2 cannot achieve better performance, as shown in Table 3. We also try to change the input raw into sRGB format. The result in the sRGB row from Table 3 shows a significant perfor-



(a) GT (b) SID [3] (c) Ours w/o CA [8] (d) Ours with CA [8]

Figure 5. Visual demonstration of checkerboard artifacts under different settings.

Table 3. Ablation study on the MCR dataset.

	DBF		DBLE	
	PSNR (dB)	SSIM	PSNR (dB)	SSIM
Baseline	**21.0607**	**0.8254**	**31.6905**	**0.9083**
Baseline wo CA [8]	20.2673	0.7948	29.2350	0.8732
Baseline wo ratio	19.8978	0.7868	29.3528	0.8878
Baseline wo packraw	20.7846	0.8034	28.8728	0.8657
Baseline l1→l2	20.4587	0.8016	30.2359	0.8974
Baseline w/o DBF	-	-	29.9946	0.8839
Baseline raw→ sRGB	18.2369	0.7625	27.3521	0.8295

mance drop, which is consistent with other works [3, 31].

The DBF module plays a key role in our system in generating the monochrome images, which assist the DBLE module in restoring the low-light images into monitor-ready sRGB images. We also explore the performance of a model without DBF module and the monochrome branch. As the results in Table 3 show, the performance drops to 29.99dB/0.883 in terms of PSNR/SSIM when the DBF module is removed, hence providing a solid validation of the DBF's effectiveness.

5. Limitations and Future Work

There are various aspects to improve in the future. The cameras we adopted in this work can only output 8-bit raw images, the 16-bit cameras will be used to collect data in the future to cover more diverse scenes and objects. Besides, the network complexity needs to be more light-weighted to deploy the proposed system in the real world. Extending the proposed work to videos will also be one future direction. We hope the work presented in this paper can provide preliminary explorations for low-light image enhancement research in community and industry. When it comes to some extremely dark images on our MCR Dataset, the existing low-light image enhancement algorithms (SID [3], LDC [31], and ours) show unsatisfactory results sometimes. The restored images usually lost the high-frequency edge information compared to the ground truth image and became blurred (see in supplementary). Extremely dark settings sometimes yield quite weak signals in each color channel, leading to those color artifacts that commonly exist in both SoTA and our methods and require further study.

6. Conclusion

Removing the Bayer-filter allows more photos to be captured by the sensor. Motivated by this fact, this work proposes an end-to-end fully convolutional network consisting of a DBF module and a dual branch low-light enhancement module to achieve low-light image enhancement on a single colored camera system. The DBF module is devised to predict the corresponding monochrome raw image from the color camera raw data input. The DBLE is designed to restore the low-light raw images based on the raw input and the DBF-predicted monochrome raw images. DBLE treats the colored raw and monochrome raw separately by using a dual branch network architecture. In the DBLE upsampling stream, features from both monochrome raw and colored raw are fused together and a channel-wise attention is applied to the fused features.

We also propose a Mono-Colored Raw paired dataset (MCR) which includes color and monochrome raw image pairs collected by a color camera with Bayer-Filter and a monochrome camera without Bayer-Filter. The dataset is collected in various scenes, and each colored raw image has a corresponding monochrome raw image captured with the same exposure settings. To better show our superiority, the SID dataset is also adopted in the evaluation. Gray image is generated from the corresponding ground truth color image in the SID dataset to serve as the monochrome image. Subsequently, a model is trained on the modified dataset to verify the performance.

Our experiments prove that significant performance can be achieved by leveraging raw sensor data and data-driven learning. Our method can overcome the checkerboard artifact which is found on U-net, while preserving the visual quality. Our quantitative experiments indicate that our methods can achieve the state-of-the-art performance: a PSNR of 31.69dB on our own dataset, and 29.65dB on the SID dataset.

References

[1] Tarik Arici, Salih Dikbas, and Yucel Altunbasak. A histogram modification framework and its application for image contrast enhancement. *IEEE Transactions on image processing*, 18(9):1921–1935, 2009. 2

[2] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3185–3194, 2019. 3

[3] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3291–3300, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[4] Xuan Dong, Guan Wang, Yi Pang, Weixin Li, Jiangtao Wen, Wei Meng, and Yao Lu. Fast efficient algorithm for enhancement of low lighting video. In *2011 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2011. 2

[5] Minhao Fan, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Integrating semantic segmentation and retinex model for low-light image enhancement. In *Proceedings of the 28th ACM International Conference on Multimedia (ACMMM)*, pages 2317–2325, 2020. 2

[6] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2511–2520, 2019. 3, 6, 7

[7] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1780–1789, 2020. 2

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. 4, 5, 6, 7, 8

[9] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7324–7333, 2019. 2, 3

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 2

[11] Guisik Kim, Dokyeong Kwon, and Junseok Kwon. Low-lightgan: Low-light enhancement via advanced generative adversarial network with task-driven training. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2811–2815. IEEE, 2019. 2

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[13] Yuma Kinoshita and Hitoshi Kiya. Fixed smooth convolutional layer for avoiding checkerboard artifacts in cnns. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3712–3716. IEEE, 2020. 5

[14] Mohit Lamba and Kaushik Mitra. Restoring extremely dark images in real time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3487–3497, 2021. 3, 6, 7

[15] Chulwoo Lee, Chul Lee, and Chang-Su Kim. Contrast enhancement based on layered difference representation of 2d histograms. *IEEE Transactions on Image Processing*, 22(12):5372–5384, 2013. 2

[16] Mading Li, Xiaolin Wu, Jiaying Liu, and Zongming Guo. Restoration of unevenly illuminated images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1118–1122. IEEE, 2018. 2

[17] Zhetong Liang, Weijian Liu, and Ruohe Yao. Contrast enhancement by nonlinear diffusion filtering. *IEEE Transactions on Image Processing*, 25(2):673–686, 2015. 2

[18] Paras Maharjan, Li Li, Zhu Li, Ning Xu, Chongyang Ma, and Yue Li. Improving extreme low-light image denoising via residual learning. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 916–921. IEEE, 2019. 3, 6, 7

[19] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016. 5

[20] Hao Ouyang, Zifan Shi, Chenyang Lei, Ka Lung Law, and Qifeng Chen. Neural camera simulators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7700–7709, 2021. 3

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted intervention*, pages 234–241. Springer, 2015. 2

[22] Eli Schwartz, Raja Giryes, and Alex M Bronstein. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE Transactions on Image Processing*, 28(2):912–923, 2018. 1, 2, 3

[23] Haonan Su and Cheolkon Jung. Low light image enhancement based on two-step noise suppression. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1977–1981. IEEE, 2017. 2

[24] Yusuke Sugawara, Sayaka Shiota, and Hitoshi Kiya. Super-resolution using convolutional neural networks without any checkerboard artifacts. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 66–70. IEEE, 2018. 5

[25] Yusuke Sugawara, Sayaka Shiota, and Hitoshi Kiya. Checkerboard artifacts free convolutional neural networks. *APSIPA Transactions on Signal and Information Processing*, 8, 2019. 5

[26] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6849–6857, 2019. 2

[27] Yuanchen Wang, Xiaonan Zhu, Yucong Zhao, Ping Wang, and Jiquan Ma. Enhancement of low-light image based on

wavelet u-net. In *Journal of Physics: Conference Series*, volume 1345, page 022030. IOP Publishing, 2019. 2

[28] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 2

[29] Xiaomeng Wu, Xinhao Liu, Kaoru Hiramatsu, and Kunio Kashino. Contrast-accumulated histogram equalization for image enhancement. In *2017 IEEE international conference on image processing (ICIP)*, pages 3190–3194. IEEE, 2017. 2

[30] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2281–2290, 2020. 2

[31] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2281–2290, 2020. 3, 6, 7, 8

[32] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50:20–29, 2019. 5

[33] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia (ACMMM)*, pages 1632–1640, 2019. 2

[34] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Eemefn: Low-light image enhancement via edge-enhanced multi-exposure fusion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13106–13113, 2020. 2