Quad Bayer Joint Demosaicing and Denoising Based on Dual Encoder Network with Joint Residual Learning

Bolun Zheng[1*], Haoran Li[1*†], Quan Chen[1 †], Tingyu Wang[1, 2], Xiaofei Zhou[1], Zhenghui Hu[3], Chenggang Yan[1]

[1]Hangzhou Dianzi University, Xiasha No.2 Street, Hangzhou, 310018, Zhejiang, China.
[2]Lishui Institute of Hangzhou Dianzi University, Semiconductor Chip Industrial Park, 323000, Zhejiang, China.
[3]Hangzhou Innovation Institute, Beihang University, Binjiang No.18 Chuanghui Street, 310018, Zhejiang, China.
{blzheng, cgyan, chenquan}@{hdu.edu.cn}, lhr970315@gmail.com, wongtyu@foxmail.com, zxforchid@outlook.com, zhenghuihu2013@163.com

Abstract

The recent imaging technology Quad Bayer color filter array (CFA) brings great imaging performance improvement from traditional Bayer CFA, but also serious challenges for demosaicing and denoising during the image signal processing (ISP) pipeline. In this paper, we propose a novel dual encoder network, namely DRNet, to achieve joint demosaicing and denoising for Quad Bayer CFA. The dual encoders are carefully designed in that one is mainly constructed by a joint residual block to jointly estimate the residuals for demosaicing and denoising separately. In contrast, the other one is started with a pixel modulation block which is specially designed to match the characteristics of Quad Bayer pattern for better feature extraction. We demonstrate the effectiveness of each proposed component through detailed ablation investigations. The comparison results on public benchmarks illustrate that our DRNet achieves an apparent performance gain (0.38dB to the second best) from the state-of-the-arts and balances performance and efficiency well. The experiments on real-world images show that the proposed method could enhance the reconstruction quality from the native ISP algorithm.

Introduction

The demands for camera capabilities have significantly increased, due to the widespread use of smartphones. However, the smartphone camera certainly suffers from the limited sensor size to achieve high-quality imaging (Ignatov, Gool, and Timofte 2020). To overcome this limitation, researchers try to turn to pixel-binning with a non-Bayer color filter array (CFA) pattern to capture more authentic scene information in a smaller sensor size, which has been proven effective for capturing images of high quality in low-light conditions (Yoo, Im, and Paik 2015). Among these non-Bayer CFA, the Quad Bayer CFA (show in Figure. 1(a)) is one of the most popular solutions.

The Quad Bayer is an extended version of the Bayer, which expands each pixel into four sub-pixels and uses a different color filter arrangement (Kim and Kim 2019). This filter array has the advantage of improving the noise performance of the image sensor by merging adjacent pixels and allowing pixels of the same color to be combined into larger pixels (Agranov et al. 2017). However, such special structure brings great challenges of accurate texture and color reconstruction, because the gap of neighbor colors is certainly enlarged. Directly applying even the SOTA demosaicing method for Bayer CFA for a Quad Bayer image would lead to a poor result (Figure. 4(b)).

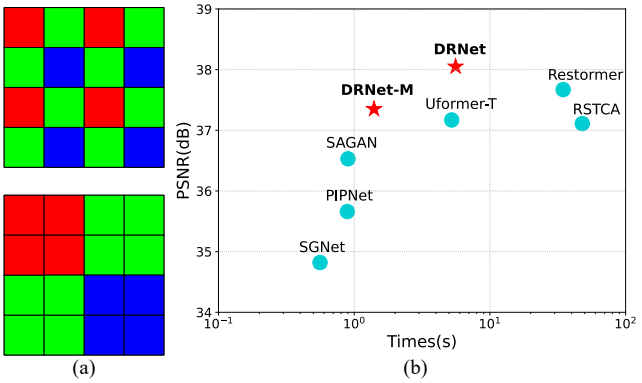


Figure 1: (a) The Bayer CFA vs. the Quad Bayer CFA. (b) PSNR vs. Running time consumption on Urban100 dataset.

The demosaicing and denoising are two key points for reconstructing high-quality RGB images from RAW images captured by Quad Bayer CFA. Directly conducting existing demosaicing for Bayer CFA on a Quad Bayer CFA image would introduce serious visual artifacts (Kim et al. 2019; Tan, Chen, and Hua 2018). Besides, the denoising is also taken into consideration in the remosacing methods to further enhance the finally reconstructed RGB images (Yang et al. 2022). However, these methods hardly ever consider the full reconstruction model against the joint demosaicing and denoising task. The additional computational cost brought by remosacing methods should also be seriously considered due to the limited computation resource of edge devices.

*These authors contributed equally.
†Corresponding author: Haoran Li and Quan Chen

Compared to remosaicing algorithms, end-to-end solutions would be more appropriate. (Ignatov, Van Gool, and Timofte 2020) proposed a pyramidal CNN architecture designed for fine-grained image restoration that implicitly learns to perform all ISP steps. (Schwartz, Giryes, and Bronstein 2018) presented an end-to-end deep learning model to tackle all of the image processing pipelines simultaneously. Though these methods get impressive performance on Byaer CFA images, ignoring the intrinsic relationship between Bayer CFA and Quad Bayer CFA makes them struggle to translate on Quad Bayer CFA images.

In this work, we propose a dual-encoder network with joint residual learning (DRNet) to overcome the limitations above towards reconstructing visually pleased RGB images from RAW images captured by a Quad Bayer CFA. Specifically, our approach leverages the dual encoders to thoroughly extract information, which includes the self-adaptive encoder (SAE) and the Quad Bayer encoder (QBE), thus resolving the difficulties in extracting information from Quad Bayer CFA. In SAE, we first decouple the joint demosaicing and denoising processes into two independent branches and involve them within one joint residual block (JRB). Then introduce a self-adaption block to cross the domain gap between the input RAW image and the reconstructed RGB image. In QBE, we design a particular Quad Bayer feature extraction block called pixel modulation block (PMB) to fully utilize the advantages of Quad Bayer CFA for noise suppressing and color restoration. To achieve a better trade-off between efficiency and performance (shown in Figure. 1), we provide a mini-version of DRNet called DRNet-M, which exhibits competitive performance and efficiency among state-of-the-art methods.

Generally, the contribution of this work can be summarized as follows:

- We propose a novel dual encoder architecture to construct an end-to-end approach for Quad Bayer joint demosaicing and denoising task.
- We decompose the joint demosaicing and denoising task into two independent sub-tasks and propose a joint residual learning block that effectively solves these two subtasks within one block.
- Against the special arrangement of the Qaud Bayer CFA, we propose a pixel modulation block to effectively Quad Bayer image feature extraction and improve the overall demosaicing and denoising performance.
- Through sufficient experiments on public benchmarks, we demonstrate our *DRNet* clearly outperforms the state-of-the-art and certainly improves the reconstruction quality from the smartphone's native ISP algorithm.

Related Work

Image demosaicing aims at reconstructing RAW images captured by digital devices into visually pleasing full-pixel RGB images, which is a fundamental step in the ISP pipeline and has been extensively studied. Research on image demosaicing can be broadly classified into two categories: traditional methods (Monno et al. 2015; Hirakawa and Parks 2005; Su 2006; Zhang and Wu 2005) and learning-based methods (Zhang et al. 2019; Kim et al. 2019; Tan, Chen, and Hua 2018; Sharif, Naqvi, and Biswas 2021; Zamir et al. 2022; Xing and Egiazarian 2022). Traditional methods typically use interpolation and filtering techniques for demosaicing. Compared to traditional methods, the learning-based methods bring significant improvement for scene adaption and degradation adaption (Zheng et al. 2020a; Hengrun et al. 2021). (Zhang et al. 2019) used residual blocks, non-local attention modules, and multi-scale feedback mechanisms to learn the local features of mosaic images. (Zamir et al. 2022) proposed two effective modules that can capture long-range pixel interactions.

During the image formation process, noise is inevitable. Merely performing demosaicing on RAW images hardly achieves satisfactory results. To address this issue, researchers concentrate on methods of joint demosaicing and denoising. Abundant research has demonstrated that such a joint process can effectively eliminate error accumulation during individual processing and improve the quality of the final reconstructed images (Liu et al. 2020; Xing and Egiazarian 2021). The joint demosaicing and denoising can be achieved through both traditional methods (Klatzer et al. 2016; Tan et al. 2017a; Condat and Mosaddegh 2012) and learning-based methods (Ehret et al. 2019; A Sharif, Naqvi, and Biswas 2021; Liu et al. 2020; Tan et al. 2017b; Xing and Egiazarian 2021). (Klatzer et al. 2016) achieved image demosaicing and denoising by minimizing the energy function in sequence.

Recently, learning-based methods exhibited great potential in image reconstruction tasks (Zheng et al. 2022; Chen et al. 2023; Zhao et al. 2021). (Ehret et al. 2019) obtained the processed demosaicing and denoising images by using a convolutional neural network with temporal and spatial redundancy information between consecutively captured images, thereby achieving the joint demosaicing and denoising task in an unsupervised way. (Liu et al. 2020) enhanced content awareness and strengthened the utilization of green channel information through the guidance of the green channel and density map. (Xing and Egiazarian 2021) proposed an end-to-end network based on residual channel attention blocks, which address image demosaicing, denoising, and super-resolution. (Zhang et al. 2022) proposed a color consistency network that can jointly store color information and enhance illumination.

In recent years, the Quad Bayer CFA has been widely used in smartphones due to its excellent performance. However, reconstructing RGB images from Quad Bayer CFA remains a challenging task (Kim et al. 2019). (A Sharif, Naqvi, and Biswas 2021) proposed a deep neural network model that includes multiple attention mechanisms, combining generative adversarial models and multiple objective functions, achieving state-of-the-art performance on joint demosaicing and denoising tasks. (Zeng et al. 2023) proposed a dual-head joint demosaicing and denoising network to convert noisy Quad Bayer CFA to noise-free Bayer CFA. (Wu et al. 2023) addressed the demosaicing and denoising processes separately in a two-stage network structure.

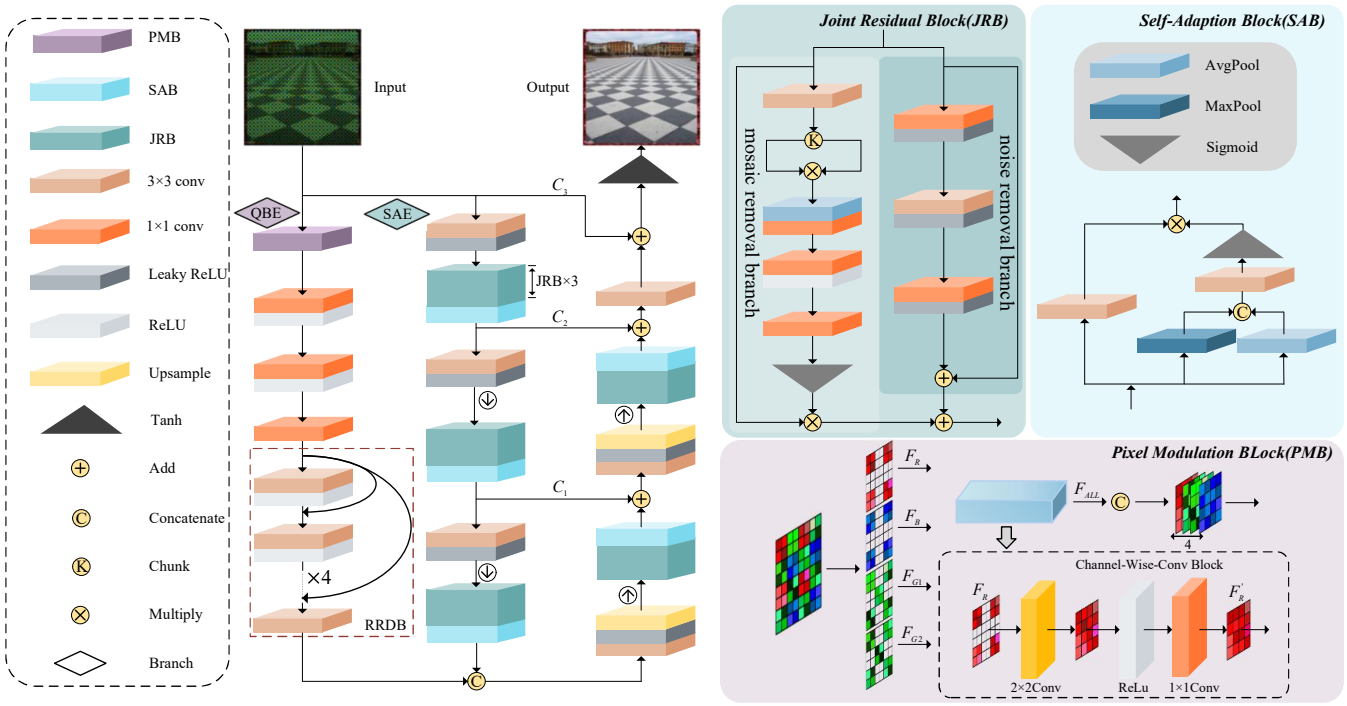


Figure 2: The overall architecture of the proposed dual encoder network (DRNet). The DRNet adopts a multi-scale encoder-decoder architecture. There are two encoders the self-adaptive encoder (SAE), and the Quad Bayer encoder (QBE), formulating the encoding part. The skip connections are introduced to connect the encoding part and decoding to produce a global residual between the RAW images and RGB images.

Proposed Method

Reconstructing a RAW image of Quad Bayer CFA can be regarded from two sights, one is remosaicing it to a common Bayer CFA image and then use existed demosaicing method to reconstruct an RGB image, while the other one is to fully regard the task as an end-to-end image restoration task. In this work, we propose a new dual encoder architecture to fuse the remoasicing and restoration within a unified framework. The two encoders in the proposed architecture are supposed to produce the Bayer CFA liked color-aware features and restored image features, respectively.

Overview of the DRNet

The architecture of the proposed DRNet is shown in Fig. 2. We adopt a widely used multiscale architecture to formulate the two encoders, the Quad Bayer encoder (QBE) and the self-adaptive encoder (SAE). The QBE starts with a pixel modulation block (PMB) to re-group the RAW pixels and produce a color-aware feature, then introduces four residual in residual dense blocks (RRDBs)(Wang et al. 2018) to enhance the features for the following fusion and upsampling. In this way, though the resolution is reduced to $\frac{1}{4}$ of the origin, most of the color and texture information is accurately extracted, which can provide sufficient guidance for the following demosaicing and denoising.

The SAE produces restored image features of three scales with joint residual learning and self-adaption. In each scale, three joint residual blocks (JRBs) and one self-adaption block (SAB) are sequentially stacked. The JRB decouples the joint demosaicing and denoising into two tasks and separately estimates the noise residuals and mosaic residuals to produce the restored image features. Meanwhile, the SAB is introduced to adapt the restored image features to the target domain for stable training.

In the decoding stage, the outputs of Quad Bayer encoder (QBE) and the top scale of self-adaptive encoder (SAE) are firstly fused, then gradually upsampled through the skip connections, and finally output the image in the sRGB domain.

Pixel Modulation

In common image sensors, the green channel accounts for half of the pixels, while the red and blue channels each account for a quarter. Quad Bayer sensors' pattern consists of two continuous uniform pixels in the vertical and horizontal directions. However, in previous work, neither of these two pieces' characteristics are simultaneously considered for joint demosaicing and denoising tasks.

To fully utilize such characteristics of the Quad Bayer pattern, we design a pixel modulation block (PMB) to achieve a color-aware feature extraction. As shown in Figure 2, the PMB first splits the input 1-channel RAW map into four sub-maps $\{F_R, F_{G1}, F_{G2}, F_B\}$. These sub-maps are then respectively sent into the channel-wise-convolution block, outputting the corresponding color-aware feature maps $\{F'_R, F'_{G1}, F'_{G2}, F'_B\}$, which are then be concatenated

together for following process.

Specifically, to convert a Quad Bayer RAW image into four sub-maps, we first regroup the adjacent 2×2 one-color pixels into a set. Assuming the left-top coordinates of a group of adjacent one-color pixels is $(2i, 2j)$, the set of this pixel group can be defined as:

$$\Phi(i,j) = \{(m,n) \,|m = i + R(0,1) \vee n = j + R(0,1)\} \tag{1}$$

where $R(0,1)$ returns a random integer between 0 and 1. Then, the whole pixel groups around the RAW images can be divided as:

$$\begin{cases} i\%2 = 0 \cup j\%4 = 0, \Phi(i,j) \subseteq R \\ i\%2 = 0 \cup j\%4 = 1, \Phi(i,j) \subseteq G_1 \\ i\%2 = 1 \cup j\%4 = 0, \Phi(i,j) \subseteq G_2 \\ i\%2 = 1 \cup j\%4 = 1, \Phi(i,j) \subseteq B \end{cases} \tag{2}$$

In this way, we can regroup the RAW images into four sub-maps, where each of them containing only one set of colors (*Red*, *Green* or *Blue*). For each sub-map, we first use a 2×2 convolution with the strides of 4×4 to fuse the adjacent 2×2 pixels and convert them to a 1-dimensional vector along the channel axis. Then a 1×1 convolution further translates the initial feature to a higher dimension feature. The above operation can be expressed as:

$$\mathbf{F'}_c = Conv_{1\times1}^{S=1}(ReLU(Conv_{2\times2}^{S=4}(\mathbf{F}_c))) \tag{3}$$

where $c \in \{R, G_1, G_2, B\}$. Along the above pipeline, the color information of adjacent pixels is fully preserved, which provides much convenience for the following noise suppression and color restoration. However, the resolution reduction and texture degradation are non-negligible. To alleviate these problems, we introduce four RRDBs after fusing the color-aware feature maps for compensation before sending them to the decoding stage.

Joint Residual Learning

The image noise and color distortion belong to different degradation models (Zheng et al. 2020b). Using classic residual-based blocks to jointly handle these two types of degradation cannot achieve satisfactory performance. To tackle this problem, we separately estimate the mosaic residual and the noise residual and propose the joint residual block (JRB). As shown in Figure. 2, JRB contains two branches, the mosaic removal branch and the noise removal branch.

Assuming the input of JRB is $\mathbf{F}_{JRB}^{in}$, a 3×3 convolution layer and a channel split layer are first introduced to produce two sub-feature map $\mathbf{F}_1$ and $\mathbf{F}_2$, which can be expressed as:

$$\mathbf{F}_1, \mathbf{F}_2 = Split(Conv(\mathbf{F}_{JRB}^{in})) \tag{4}$$

where $Conv(\cdot)$ and $Split(\cdot)$ denotes the 3×3 convolution and the channel split. Notice that the $\mathbf{F}_1$ and $\mathbf{F}_2$ get the same shape. Then we generate a basic color error map $\mathbf{S}_{basic}$ as:

$$\mathbf{S}_{basic} = P_{avg}(\mathbf{F}_1 \cdot \mathbf{F}_2) \tag{5}$$

where $P_{avg}(\cdot)$ denotes a global average pooling along the channel axis. Then we can obtain the final color error map $\mathbf{S}$ by sequentially introducing a 1×1 convolution with ReLU (Glorot, Bordes, and Bengio 2011) activation and another 1×1 convolution with Sigmoid (Glorot, Bordes, and Bengio 2011) activation. Finally, the mosaic residual $\mathbf{R}_{mosaic}$ is calculated by:

$$\mathbf{R}_{mosaic} = \mathbf{S} \cdot \mathbf{F}_{JRB}^{in} \tag{6}$$

In the noise removal branch, we simply use a 1×1 convolution, a 3×3 convolution, and another 1×1 to obtain the noise residual R_{noise} as the noise is additive. Specifically, these three convolution layers all accompany a Leaky ReLU (Chung et al. 2014) activation. Then, we can have the output of the JRB as:

$$\mathbf{F}_{JRB}^{out} = \mathbf{F}_{JRB}^{in} + \mathbf{R}_{mosaic} + \mathbf{F}_{noise} \tag{7}$$

where $\mathbf{F}_{JRB}^{out}$ denotes the output of the JRB.

Self Adaption

It's clear that the source RAW domain and the target sRGB domain are totally different. Therefore, it's necessary to introduce a domain adaption block to cross the domain gap between the source and target.

To this end, we introduce the self-adaption block (SAB) to achieve a self-adaptive domain adaption. Assuming the input of the SAB is denoted as $\mathbf{F}_{SAB}^{in}$, we first use a global max pooling layer and a global average pooling layer to distil the information from the $\mathbf{F}_{SAB}^{in}$ along the channel axis, which can be expressed as:

$$\mathbf{F}_{max} = P_{max}(\mathbf{F}_{SAB}^{in}) \tag{8}$$

$$\mathbf{F}_{avg} = P_{avg}(\mathbf{F}_{SAB}^{in}) \tag{9}$$

where $P_{max}(\cdot)$ denotes the global max pooling along the channel axis. Then we concatenate the $\mathbf{F}_{max}$ and $\mathbf{F}_{avg}$ along the channel axis and calculate the adaption map $\mathbf{A}$ as:

$$\mathbf{A} = Sigmoid(Conv(< \mathbf{F}_{max}, \mathbf{F}_{avg} >)) \tag{10}$$

where $<,>$ denotes the concatenate operation. So that we can now achieve the domain adaption and produce the output of the SAB as:

$$\mathbf{F}_{SAB}^{out} = \mathbf{A} \cdot Conv(\mathbf{F}_{SAB}^{in}) \tag{11}$$

where $\mathbf{F}_{SAB}^{out}$ denotes the output of the SAB.

Loss Function

To train the proposed DRNet, we adopt a hybrid loss function formulated by adversarial loss, color reconstruction loss, perceptual loss, and L1 loss, which can be defined as:

$$\mathcal{L} = \lambda_D \mathcal{L}_D + \mathcal{L}_1 + \mathcal{L}_P + \mathcal{L}_C \tag{12}$$

where $\mathcal{L}_D$ denotes the adversarial loss and $\lambda_D = 10^{-4}$ is a hyperparameter, $mathcal L_1$ denotes the L1 loss, $\mathcal{L}_P$ denotes the perceptual loss and $\mathcal{L}_C$ denotes the color reconstruction loss. Specifically, the perceptual loss is constructed with the VGG-19 backbone, which the $\mathcal{L}_C$ is to measure the color difference (ΔE) defined in CIEDE2000(Luo, Cui, and Rigg 2001), which is written as:

$$\mathcal{L}_c(X, Y) = \triangle E(X, Y) \tag{13}$$

where X and Y denote the outputs of DRNet and the corresponding groundtruth.

Method	Params (M)	σ	Urban100	BSD100	MCM	KODAK	Set14	Average
			PSNR/SSIM					
SGNet	0.912	5	34.82/0.9634	38.35/0.9786	32.42/0.9025	35.63/0.9616	33.53/0.9301	34.95/0.9472
		15	33.58/0.9466	35.41/0.9488	30.85/0.8691	33.25/0.9262	31.65/0.9014	32.95/0.9184
		25	31.05/0.9137	32.89/0.9154	28.90/0.8242	31.19/0.8869	30.04/0.8682	30.81/0.8816
PIPNet	3.459	5	35.66/0.9670	39.43/0.9803	35.55/0.9470	37.00/0.9672	35.47/0.9414	36.62/0.9605
		15	33.85/0.9481	36.68/0.9586	33.92/0.9234	34.68/0.9394	33.59/0.9198	34.54/0.9378
		25	32.28/0.9285	34.62/0.9353	32.45/0.8990	32.53/0.9046	32.12/0.8977	32.80/0.9129
Uformer	8.203	5	37.17/0.9788	39.76/0.9825	34.61/0.9365	36.17/0.9636	34.81/0.9385	36.50/0.9599
		15	34.57/0.9547	36.33/0.9525	33.24/0.9125	33.88/0.9286	32.90/0.9122	34.18/0.9321
		25	32.27/0.9252	34.02/0.9227	31.87/0.8824	32.02/0.8898	31.34/0.8803	32.30/0.9001
PyNet	28.937	5	35.79/0.9714	38.35/0.9786	32.42/0.9025	35.63/0.9616	33.53/0.9301	35.14/0.9488
		15	33.58/0.9466	35.41/0.9488	30.85/0.8691	33.25/0.9262	31.65/0.9014	32.95/0.9184
		25	31.05/0.9137	32.89/0.9154	28.90/0.8242	31.19/0.8869	30.04/0.8682	30.81/0.8816
SAGAN	22.557	5	36.53/0.9767	39.93/0.9825	34.28/0.9358	35.67/0.9626	34.64/0.9382	36.21/0.9591
		15	34.31/0.9533	36.61/0.9541	32.96/0.9087	33.69/0.9278	32.90/0.9123	34.09/0.9312
		25	31.95/0.9222	33.91/0.9206	31.45/0.8746	31.73/0.8861	31.21/0.8805	32.05/0.8967
Restormer	26.097	5	37.67/0.9777	39.39/0.9760	35.88/0.9493	36.72/0.9618	35.55/0.9421	37.04/0.9614
		15	34.79/0.9556	36.18/0.9498	34.25/0.9286	34.45/0.9339	33.59/0.9201	34.65/0.9376
		25	32.84/0.9353	34.31/0.9285	32.90/0.9078	32.79/0.9074	32.21/0.8996	33.01/0.9157
RSTCA	0.921	5	37.11/0.9800	38.29/0.9820	34.83/0.9387	36.19/0.9661	34.92/0.9425	36.27/0.9619
		15	34.67/0.9584	35.41/0.9549	33.50/0.9165	34.13/0.9360	33.16/0.9189	34.17/0.9368
		25	32.57/0.9331	33.28/0.9250	32.20/0.8921	32.40/0.9029	31.74/0.8930	32.44/0.9092
DRNet-M (Ours)	1.405	5	37.35/0.9810	39.99/0.9827	35.22/0.9455	36.39/0.9660	35.22/0.9442	36.83/0.9638
		15	34.84/0.9580	36.77/0.9563	33.77/0.9219	34.28/0.9366	33.41/0.9200	34.61/0.9386
		25	32.72/0.9334	34.57/0.9309	32.32/0.8943	32.53/0.9053	31.94/0.8948	32.82/0.9116
DRNet (Ours)	5.595	5	**38.05/0.9837**	**40.48/0.9849**	**36.01/0.9520**	**37.05/0.9685**	**35.73/0.9467**	**37.46/0.9670**
		15	**35.37/0.9599**	**37.07/0.9581**	**34.41/0.9301**	**34.96/0.9399**	**33.91/0.9268**	**35.14/0.9428**
		25	**32.95/0.9355**	**34.86/0.9342**	**32.99/0.9084**	**32.96/0.9110**	**32.32/0.9006**	**33.22/0.9178**

Table 1: Performance comparison of state-of-the-art methods for joint demosaicing and denoising on Quad Bayer CFA. The BOLD and UNDERLINE indicates the best and second best results respectively.

Experiments

In this section, we first introduce the implementation details of the model settings, training details, and related datasets. Then, the comparison of public benchmarks among the proposed method and several state-of-the-art methods will be conducted to demonstrate the superiority of our DRNet. Moreover, a set of ablation experiments focusing on the key components and major contributions, as we argued before, will be included to illustrate the effectiveness and importance of each of them.

Implementation Details

We formulate the proposed DRNet by setting the number of feature map channels as [64, 128, 256] for each scale in SAE. Because the output of QBE exists on the top scale we let its output get 128 channels. The mini-version of the DRNet, namely DRNet-M, gets the feature maps in the self-adaptive encoder (SAE) of [32, 64, 128] channels for three scales, while the Quad Bayer encoder (QBE) outputs a 64-channel feature map. To train the proposed method, we re-group the training data into a set of 128×128 sized patches with a batchsize of 16. We adopt the Adam optimizer(Kingma and Ba 2014) with the initialized learning rate of 3×10^{-4}. The learning rate will smoothly decrease to 3×10^{-5} during the training stage of a total 100 epochs. All the experiments are conducted with the Pytorch framework on a Nvidia RTX3090 GPU server.

We utilize DIV2K (Agustsson and Timofte 2017) and Flickr2K (Timofte et al. 2017) as training set, while the Urban100 (Cordts et al. 2016), BSD100 (Martin et al. 2001), MCM (Woo et al. 2018), Kodak (Loui et al. 2007), and Set14 are selected as testing set. To simulate the noise Quad Bayer CFA image, we re-sample the RGB image according to the Quad Bayer CFA and add Gaussian noise of $\sigma = 5/15/25$ to synthesize the input. Moreover, a real-world dataset SIDD (Abdelhamed, Lin, and Brown 2018) is int roduced to further study the performance of compared methods for real-world applications. In this case, the additional noise won't be included when synthesising the inputs.

Comparison to State-of-the-Arts

To demonstrate the superiority of our DRNet, we conduct a fair comparison to several state-of-the-art methods, including two demosaicing methods Pynet (Ignatov, Van Gool, and Timofte 2020) and Rstca (Xing and Egiazarian 2022), two joint demosaicing and denoising methods SGNet (Liu et al. 2020) and PIPNet (A Sharif, Naqvi, and Biswas 2021), as well as two image restoration methods Ufomrer (Wang et al. 2022) and Restormer (Zamir et al. 2022). We mainly investigate the PSNR and SSIM (Wang et al. 2004) indexes to

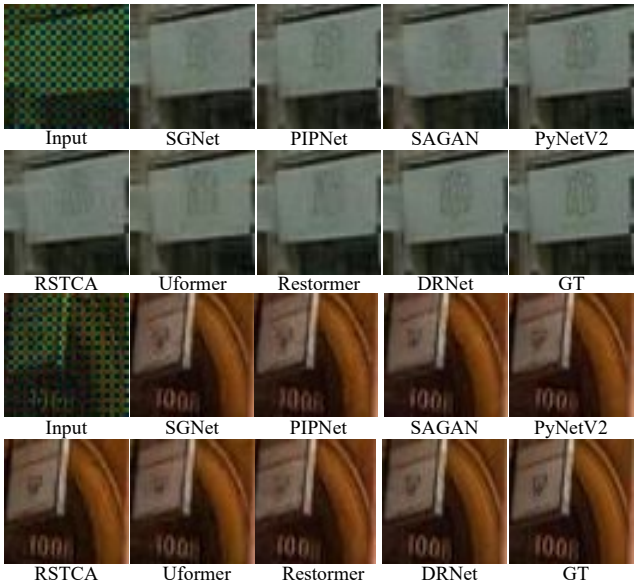


Figure 3: The visualized results of all compared methods for joint demosaicing and denoising on Quad Bayer CFA.

Methods	Inference Time(s)				PSNR
	r=192	r=256	r=384	r=512	σ=5
PIPNet	0.084	0.141	0.293	0.519	36.62
SAGAN	0.183	0.301	0.599	1.073	36.21
Restormer	0.658	1.107	✗	✗	37.04
RSTCA	0.078	✗	✗	✗	36.27
DRNet-M	**0.049**	**0.079**	**0.149**	**0.260**	36.83
DRNet	0.101	0.178	0.375	0.665	**37.46**

Table 2: The comparisons for inference time with the $r \times r$ sized patch input on the mobile processor and the average PSNR. The ✗ denotes the out of memory.

illustrate the performance of all compared methods.

We provide the quantitative comparison results for all compared methods in Table. 1. As shown, the proposed DRNet outperforms all the other methods on all datasets and performance indexes and beats the second-best method by 0.42/0.49/0.21dB in terms of PSNR for $\sigma = 5/15/20$. Moreover, the mini-model DRNet-M achieves the lowest MACs and still keeps competitive performance that gets the second-best average PSNR and SSIM performance. These observations certainly demonstrate the superiority of the performance and efficiency of the proposed method. We also provide the visualized details of all compared methods in Figure. 3. Benefiting from the dual-encoder architecture, our DRNet could well preserve the color texture details and successfully suppress the noise. At the same time, the other methods suffer from the blurring effects and noise contamination, especially around the areas of strong edges and great color changes.

Experiments on the Edge Device We conduct the ablation experiments by comparing models constructed and trained with different combinations of proposed components. Table.2 reports the inference time of several methods on the "Snapdragon 8 Gen 2" mobile processor. The transformer-based methods Restormer and RSTCA fail to process images of larger sizes due to the limited memory. Besides, our DRNet-M exhibits a good trade-off between efficiency and effectiveness that gets the best performance (except for Restormer) with the best efficiency.

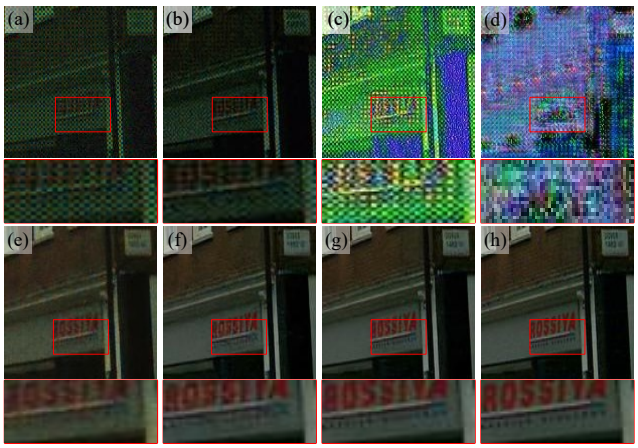


Figure 4: The visualized results of partially removing each component from the complete model ($\sigma = 25$). (a) input. (b) PIPNet for Bayer CFA. (c) removing the mosaic removal branch. (d) removing the self-adaption block (SAB). (e) removing the noise removal branch. (f) removing the pixel modulation block (PMB). (g) DRNet. (h) GT.

Modules		Urban100	BSD100
NRB	MRB	PSNR/SSIM/LPIPS	PSNR/SSIM/LPIPS
		26.52/0.8405/0.169	28.57/0.8302/0.251
	✓	31.24/0.9035/0.066	32.84/0.8831/0.116
✓		29.57/0.8880/0.111	30.29/0.8463/0.224
✓	✓	**32.95/0.9355/0.039**	**34.86/0.9342/0.063**

Table 3: Experiment results for the noise removal branch (denoted as NRB in the table) and the mosaic removal branch (denoted as MRB in the table) on Urban100 and BSD100 datasets. (σ=25)

Modules Verification

As we argued before, each of the proposed components gets a clear target for the joint demosaicing and denoising task. To demonstrate these arguments, we take a fully trained DRNet as a baseline and visualize the output of partially removing each component from the complete model. In this way, we can easily distinguish the role that each component plays to finally achieve the joint demosaicing and denoising.

Modules in QBE Figure. 4 shows the results by partially removing the QBE, the noise removal branch, and the mosaic removal branch of JRB and the SAB. In the QBE, we design PMB to regroup RAW pixels to generate color-aware and texture information features. Removing PMB (Figure.4 (f)) leads to the texture recovery markedly deteriorated com-

pared to our DRNet (Figure.4 (g)).

Modules in SAE In the self-adaptive encoder (SAE), JRB is the key for joint demosaicing and denoising. Image denoising is the process of removing additive noise while demosaicking can be defined as tone mapping that restores image colors through dot multiplication (Zheng et al. 2021). Removing SAB makes the reconstructed images unreadable (Figure.4 (d)), which illustrates that the SAB is a key point for domain transfer.

Branches in JRB There are two branches, mosaic removal branch and noise branch in the JRB. Removing the mosaic removal branch leads to color distortion (Figure.4 (c)). After removing the noise removal branch (Figure.4 (e)), the noise significantly increases. Besides, we also provide quantitative ablation experiment in Table. 3 that removing any of branches in the JRB would certainly lead considerable performance degradation. Thus, all the components work as we argue in the previous sections.

Modules		Urban100	BSD100
PMB	RRDB	PSNR/SSIM/LPIPS	PSNR/SSIM/LPIPS
		29.28/0.8713/0.042	29.25/0.8916/0.074
	✓	32.52/0.9306/0.045	33.90/0.9278/0.067
✓		32.68/0.9176/0.042	34.17/0.9240/0.073
✓	✓	**32.95/0.9355/0.039**	**34.86/0.9342/0.063**

Table 4: Ablation experiment results for PMB and RRDB on Urban100 and BSD100 datasets.(σ=25)

Ablation Investigation

In this subsection, we provide sets of experiments to demonstrate the effectiveness of each component in our DRNet.

Ablation Experiments for QBE We first investigate the effectiveness of modules pixel modulation block (PMB) and RRDB in QBE. The results are shown in Table. 4, removing the RRDB leads to an inevitable performance reduction while removing PMB (replacing PMB with a pixelshuffle (Shi et al. 2016) layer) leads to an even worse situation. While if we remove both PMB and RRDBs, the performance is sharply reduced. This observation demonstrates that the missing of QBE is unacceptable. In QBE, the PMB plays a more important role than RRDBs in Quad Bayer feature extraction, while the feature enhancement provided by RRDBs also should not be ignored.

Modules		Urban100	BSD100
JRB	SAB	PSNR/SSIM/LPIPS	PSNR/SSIM/LPIPS
		31.89/0.9163/0.065	34.31/0.9220/0.074
✓		29.62/0.8801/0.094	31.13/0.8188/0.279
	✓	31.24/0.9126/0.047	31.22/0.8970/0.120
✓	✓	**32.95/0.9355/0.039**	**34.86/0.9342/0.063**

Table 5: Ablation experiment results for JRB and SAB on Urban100 and BSD100 datasets.(σ=25)

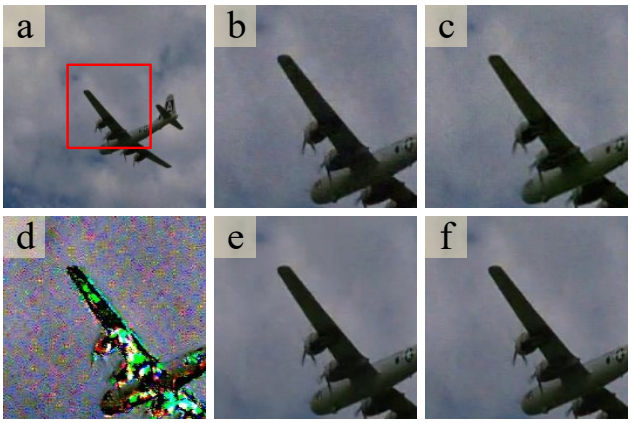


Figure 5: The visualized results of ablation experiments for JRB and SAB. (a) GT. (b) removing SAB and replacing JRBs with resblocks. (c) replacing JRBs with resblocks. (d) removing SAB. (e) reconstructed with DRNet. (f) original.

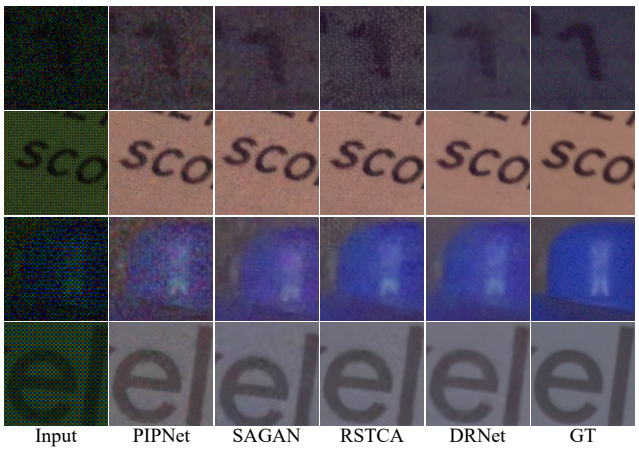
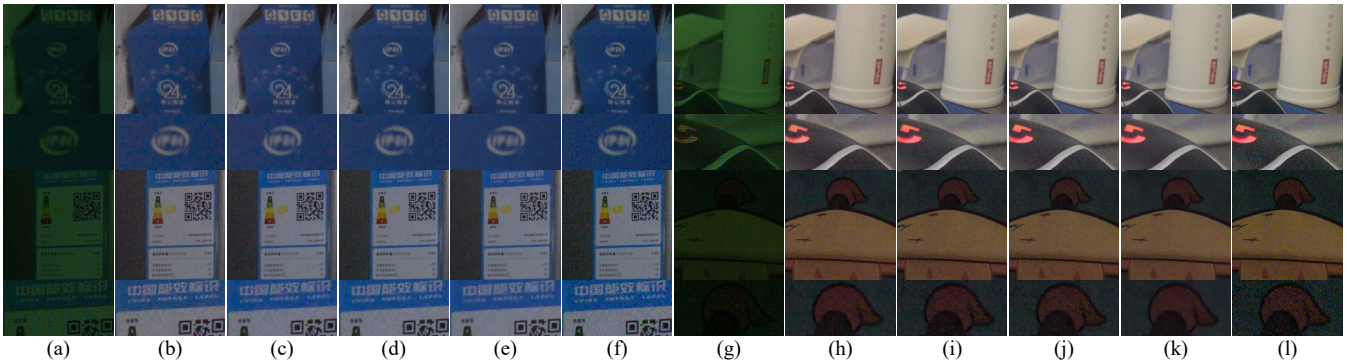


Figure 6: Visualized results of Quad Bayer joint demosaicing and denoising on SIDD dataset.

Ablation Experiments for SAE Then we investigate the effectiveness of components in SAE, including the joint residual block and the self-adaption block, by training the models that partially remove one or both of them. Specifically, to fairly evaluate the effectiveness of JRB, we remove JRBs by replacing them with common residual blocks (Lim et al. 2017) of closed computation cost. As shown in Table. 5 and Figure. 5, the network fails to give accurate reconstruction results without the domain adaption provided by SAB. We hold the opinion that the JRB and SAB should work in concert for the best reconstruction effects. Without the domain adaption provided by SAB, it's difficult for JRB to learn an accurate residual estimation. Besides, the JRB should be the most valuable part for the joint demosaicing and denoising as it provides an explicit degradation model for jointly estimating the residuals of the noise and mosaic artifacts. By contrast, the common residual blocks cannot well handle the demosaicing and denoising tasks at



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Figure 7: The visualized results on real data. (a)&(g) Input. (b)&(h) PIPNet. (c)&(i) SAGAN. (d)&(j) RSTCA. (e)&(k) DRNet. (f)&(l) smartphone's native ISP algorithm.

Method	Time(s)	SIDD		
		PSNR	SSIM	LPIPS
PIPNet	9.48	28.95	0.7704	0.340
SAGAN	9.92	34.31	0.8377	0.209
RSTCA	19.86	35.25	0.8502	0.192
DRNet	12.58	**35.72**	**0.8621**	**0.177**

Table 6: Comparison results of Quad Bayer joint demosaicing and denoising on SIDD datasets.

the same time, though they take similar computation costs. However, there is an accident that removing both JRB and SAB leads to even better results than singly removing one of them. This observation again indicates the importance of simultaneously introducing JRB and SAB because the SAB provides the necessary domain adaption to ensure the JRB can be well trained.

Generalization to Real Data

To further demonstrate the superiority, we conduct an additional experiment on real-world data involving the SOTA methods exhibited in Table. 6 including PIPNet (A Sharif, Naqvi, and Biswas 2021), SAGAN (Sharif, Naqvi, and Biswas 2021) and RSTCA (Xing and Egiazarian 2022). Considering the real-world noise is quite different from the Gaussian noise, we adopt a real-world denoising dataset SIDD (Abdelhamed, Lin, and Brown 2018) to train all compared models for a fair comparison. We adopt the same pipeline described in Sec. 4.1 to construct inputs by converting the noise RGB images to Quad Bayer format.

The quantitative and qualitative results of compared methods on the SIDD dataset are available in Table. 6 and Figure. 6. The performance gaps among the compared methods are further enlarged on the real-world noise images. The PIPNet suffers from non-uniformly and irregularly distributed noise along with the demosaicing process. By contrast, decomposing the joint demosaicing and denoising task into two independent sub-tasks clearly promotes robustness and effectiveness against complex real-world images.

Then, we test these methods with their models trained on the SIDD dataset on authentic Quad Bayer images captured by a smartphone. We adopt the OnePlus9R smartphone, whose camera sensor is a Quad Bayer CFA, to capture testing images. Figure. 7 shows the results of the compared methods and the smartphone's native ISP algorithm. From the visualized results, reconstructing RGB images from Quad Bayer images remains a great challenge for native ISP algorithms. Among these compared methods, our DRNet exhibits the best performance for suppressing the real-wolrd noise and preserving the details (Figure. 7 (e)&(k)).

Conclusion

In this paper, we propose a novel deep learning approach *DRNet* for joint demosaicing and denoising for Quad Bayer CFA. The proposed DRNet employs a dual encoder architecture including a Quad Bayer encoder (QBE) and self-adaptive encoder (SAE) to jointly fuse the Quad Bayer features and restored image features. In the QBE, we propose a pixel modulation block to mostly preserve the color information and introduce RRDBs to compensate for the lost texture details. In the SAE, we propose a joint residual learning block to accomplish the denoising and demosaicing tasks within one residual block and introduce a self-adaption block to cross the domain gap between RAW inputs and sRGB outputs. We demonstrate the effectiveness and superiority of our DRNet through sufficient experiments on multiple benchmarks and real-world evaluations.

Acknowledgements

This work is supported by the National Nature Science Foundation of China No. 62001146, U21B2024, the Key R&D Program of Zhejiang under Grant No. 2023C01044. This work is also suppoted by the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grants No. GK239909299001-013 and GK229909299001-009.

References

A Sharif, S.; Naqvi, R. A.; and Biswas, M. 2021. Beyond joint demosaicking and denoising: An image processing pipeline for a pixel-bin image sensor. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 233–242.

Abdelhamed, A.; Lin, S.; and Brown, M. S. 2018. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1692–1700.

Agranov, G. A.; Molgaard, C.; Bahukhandi, A.; Lee, C.; and Li, X. 2017. Pixel binning in an image sensor. *US Patent*, 9: B2.

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 126–135.

Chen, R.; Zheng, B.; Zhang, H.; Chen, Q.; Yan, C.; Slabaugh, G.; and Yuan, S. 2023. Improving dynamic HDR imaging with fusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 340–349.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Condat, L.; and Mosaddegh, S. 2012. Joint demosaicking and denoising by total variation minimization. In *2012 19th IEEE International Conference on Image Processing*, 2781–2784. IEEE.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Ehret, T.; Davy, A.; Arias, P.; and Facciolo, G. 2019. Joint demosaicking and denoising by fine-tuning of bursts of raw images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8868–8877.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323. JMLR Workshop and Conference Proceedings.

Hengrun, Z.; Bolun, Z.; Shanxin, Y.; Hua, Z.; Chenggang, Y.; Liang, L.; and Gregory, S. 2021. CBREN: Convolutional Neural Networks for Constant Bit Rate Video Quality Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*.

Hirakawa, K.; and Parks, T. W. 2005. Adaptive homogeneity-directed demosaicing algorithm. *Ieee transactions on image processing*, 14(3): 360–369.

Ignatov, A.; Gool, L.; and Timofte, R. 2020. Replacing Mobile Camera ISP with a Single Deep Learning Model.

Ignatov, A.; Van Gool, L.; and Timofte, R. 2020. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 536–537.

Kim, I.; Song, S.; Chang, S.; Lim, S.; and Guo, K. 2019. Deep image demosaicing for submicron image sensors. *Journal of Imaging Science and Technology*, 63(6): 60410–1.

Kim, Y.; and Kim, Y. 2019. High-Sensitivity Pixels with a Quad-WRGB Color Filter and Spatial Deep-Trench Isolation. *Sensors*, 19(21): 4653.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization.

Klatzer, T.; Hammernik, K.; Knobelreiter, P.; and Pock, T. 2016. Learning joint demosaicing and denoising based on sequential energy minimization. In *2016 IEEE International Conference on Computational Photography (ICCP)*, 1–11. IEEE.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 136–144.

Liu, L.; Jia, X.; Liu, J.; and Tian, Q. 2020. Joint demosaicing and denoising with self guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2240–2249.

Loui, A.; Luo, J.; Chang, S.-F.; Ellis, D.; Jiang, W.; Kennedy, L.; Lee, K.; and Yanagawa, A. 2007. Kodak's consumer video benchmark data set: concept definition and annotation. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, 245–254.

Luo, M. R.; Cui, G.; and Rigg, B. 2001. The development of the CIE 2000 colour-difference formula: CIEDE2000. *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*, 26(5): 340–350.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, 416–423. IEEE.

Monno, Y.; Kiku, D.; Tanaka, M.; and Okutomi, M. 2015. Adaptive residual interpolation for color image demosaicking. In *2015 IEEE International Conference on Image Processing (ICIP)*, 3861–3865. IEEE.

Schwartz, E.; Giryes, R.; and Bronstein, A. M. 2018. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE Transactions on Image Processing*, 28(2): 912–923.

Sharif, S.; Naqvi, R. A.; and Biswas, M. 2021. SAGAN: Adversarial spatial-asymmetric attention for noisy NONA-bayer reconstruction. *arXiv preprint arXiv:2110.08619*.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.

Su, C.-Y. 2006. Highly effective iterative demosaicing using weighted-edge and color-difference interpolations. *IEEE Transactions on Consumer Electronics*, 52(2): 639–645.

Tan, D. S.; Chen, W.-Y.; and Hua, K.-L. 2018. DeepDemosaicking: Adaptive image demosaicking via multiple deep fully convolutional networks. *IEEE Transactions on Image Processing*, 27(5): 2408–2419.

Tan, H.; Zeng, X.; Lai, S.; Liu, Y.; and Zhang, M. 2017a. Joint demosaicing and denoising of noisy bayer images with ADMM. In *2017 IEEE International Conference on Image Processing (ICIP)*, 2951–2955. IEEE.

Tan, R.; Zhang, K.; Zuo, W.; and Zhang, L. 2017b. Color image demosaicking via deep residual learning. In *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, volume 2, 6.

Timofte, R.; Agustsson, E.; Van Gool, L.; Yang, M.-H.; and Zhang, L. 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 114–125.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, 0–0.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17683–17693.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wu, X.; Fan, Z.; Zheng, J.; Wu, Y.; and Zhang, F. 2023. Learning to Joint Remosaic and Denoise in Quad Bayer CFA via Universal Multi-scale Channel Attention Network. In Karlinsky, L.; Michaeli, T.; and Nishino, K., eds., *Computer Vision – ECCV 2022 Workshops*, 147–160. Cham: Springer Nature Switzerland. ISBN 978-3-031-25072-9.

Xing, W.; and Egiazarian, K. 2021. End-to-end learning for joint image demosaicing, denoising and super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3507–3516.

Xing, W.; and Egiazarian, K. 2022. Residual swin transformer channel attention network for image demosaicing. In *2022 10th European Workshop on Visual Information Processing (EUVIP)*, 1–6. IEEE.

Yang, Q.; Yang, G.; Jiang, J.; Li, C.; Feng, R.; Zhou, S.; Sun, W.; Zhu, Q.; Loy, C. C.; and Gu, J. 2022. MIPI 2022 Challenge on Quad-Bayer Re-mosaic: Dataset and Report. In *ECCV Workshop*.

Yoo, Y.; Im, J.; and Paik, J. 2015. Low-light image enhancement using adaptive digital pixel binning. *Sensors*, 15(7): 14917–14931.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5728–5739.

Zeng, H.; Feng, K.; Cao, J.; Huang, S.; Zhao, Y.; Luong, H.; Aelterman, J.; and Philips, W. 2023. Inheriting Bayer's Legacy-Joint Remosaicing and Denoising for Quad Bayer Image Sensor.

Zhang, L.; and Wu, X. 2005. Color demosaicking via directional linear minimum mean square-error estimation. *IEEE Transactions on Image Processing*, 14(12): 2167–2178.

Zhang, Y.; Li, K.; Li, K.; Zhong, B.; and Fu, Y. 2019. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*.

Zhang, Z.; Zheng, H.; Hong, R.; Xu, M.; Yan, S.; and Wang, M. 2022. Deep Color Consistent Network for Low-Light Image Enhancement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1889–1898.

Zhao, H.; Zheng, B.; Yuan, S.; Zhang, H.; Yan, C.; Li, L.; and Slabaugh, G. 2021. CBREN: Convolutional neural networks for constant bit rate video quality enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4138–4149.

Zheng, B.; Chen, Q.; Yuan, S.; Zhou, X.; Zhang, H.; Zhang, J.; Yan, C.; and Slabaugh, G. 2022. Constrained Predictive Filters for Single Image Bokeh Rendering. *IEEE Transactions on Computational Imaging*, 8: 346–357.

Zheng, B.; Chen, Y.; Tian, X.; Zhou, F.; and Liu, X. 2020a. Implicit dual-domain convolutional network for robust color image compression artifact reduction. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11): 3982–3994.

Zheng, B.; Yuan, S.; Slabaugh, G.; and Leonardis, A. 2020b. Image demoireing with learnable bandpass filters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3636–3645.

Zheng, B.; Yuan, S.; Yan, C.; Tian, X.; Zhang, J.; Sun, Y.; Liu, L.; Leonardis, A.; and Slabaugh, G. 2021. Learning Frequency Domain Priors for Image Demoireing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.