PRNet: Pyramid Restoration Network for RAW Image Super-Resolution

Mingyang Ling, Kan Chang, *Member, IEEE,* Mengyuan Huang, Hengxin Li, Shuping Dang, *Senior Member, IEEE,* and Baoxin Li, *Senior Member, IEEE*

***Abstract*—Typically, image super-resolution (SR) methods are applied to the standard RGB (sRGB) images produced by the image signal processing (ISP) pipeline of digital cameras. However, due to error accumulation, low bit depth and the nonlinearity with scene radiance in sRGB images, performing SR on them is sub-optimal. To address this issue, a RAW image SR method called pyramid restoration network (PRNet) is proposed in this paper. Firstly, PRNet takes the low-resolution (LR) RAW image as input, and generates a rough estimation of the SR result in the linear color space. Afterwards, a pyramid refinement (PR) sub-network refines image details in the intermediate SR result and corrects its colors in a divide-and-conquer manner. To learn the appropriate colors for displaying, external guidance is extracted from the LR reference image in the sRGB color space, and then fed to the PR sub-network. To effectively incorporate the external guidance, the cross-layer correction module (CLCM), which fully investigates the long-range interactions between two input features, is introduced in the PR sub-network. Moreover, as different frequency components decomposed from the same image are highly correlated, in the PR sub-network, the refined features from a lower layer are utilized to support the feature refinement in an upper layer. Extensive experiments presented in this paper demonstrate that the proposed method is capable of recovering fine details and small structures in images while producing vivid colors that align with the output of a specific camera ISP pipeline.**

***Index Terms*—Super-resolution, color image demosaicking, RAW image processing, image signal processing pipeline, Laplacian pyramid decomposition.**

I. Introduction

IMAGE super-resolution (SR) aims to restore high-resolution (HR) images from given low-resolution (LR) images. Over the past few years, numerous convolutional neural network (CNN)-based SR methods have been proposed [1]–[3] and achieved remarkable performance. However, usually, the SR methods are applied to the standard RGB (sRGB) images produced by the image signal processing (ISP) pipeline of digital cameras [4], thus leading to the following three main drawbacks:

1) *Error accumulation*: Image restoration operations, including color demosaicking [5]–[9] and noise reduction [10]–[12] (and sometimes deblurring), have been applied in the ISP pipeline [13]. Moreover, sRGB images are usually compressed and stored, e.g., in JPEG format, and thus inevitably brings quantization noises and compression artifacts. Since these operations are processed separately, the errors produced by each operation may accumulate and spread within sRGB images, potentially affecting the SR method.
2) *Low bit depth*: Typically, in sRGB images, only 8 bits are used to record one color channel of a single pixel. Compared to a higher bit depth, such as 12 or 14 bits, the relatively low bit depth used in sRGB images limits the recorded visual information and thus considerably restricts the quality of SR results.
3) *Nonlinearity with scene radiance*: In a common ISP pipeline, nonlinear operations, such as tone curve, look-up table (LUT) and gamma transformation, are applied to convert the colors from the linear color space to the nonlinear sRGB space [14], [15]. As a result, the pixel intensity in sRGB images is no longer linear to the scene radiance [16], which makes the SR task even more challenging.

Fortunately, in digital cameras, along with the compressed, low-bit-depth sRGB images, RAW images are also available. Since the RAW images have not been processed by the ISP pipeline and contain a higher bit depth (usually 12 or 14 bits), directly conducting RAW image SR is an appealing approach. It is worth noting that, due to the usage of color filter array (CFA), only one color channel is recorded for each pixel in a RAW image. Consequently, it is necessary to restore the other two missing color channels while enhancing the spatial resolution of an image. To this end, recently, some joint demosaicking and SR (JDSR) approaches have been proposed [13], [17]–[20].

Besides demosaicking, another important issue that needs to be addressed is the establishment of a mapping from the RAW image color space to the sRGB space for ensuring proper color rendering. However, the specific system design and parameters of ISP pipeline are typically proprietary and closely guarded by camera manufacturers, making the real ISP pipeline practically a "black box" [15]. Moreover, different brands and types of cameras have distinct settings, leading to their unique camera styles. As a result, without the prior

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62171145, and also by Guangxi Key R&D Program under Grant AB23075106. *(Corresponding author: Kan Chang.)*

Mingyang Ling is with the School of Electrical Engineering, Guangxi University, Nanning 530004, China (e-mail: lingmy@st.gxu.edu.cn).

Kan Chang and Mengyuan Huang are with the School of Computer and Electronic Information, Guangxi University, Nanning 530004, China, and also with the Guangxi Key Laboratory of Multimedia Communications and Network Technology, Guangxi University, Nanning 530004, China (e-mail: kanchang@gxu.edu.cn; huangmengyuan@st.gxu.edu.cn).

Hengxin Li is with Shenzhen Xiaomi Communications Co., Ltd, Shenzhen 518000, China (e-mail: lihengxin@xiaomi.com).

Shuping Dang is with the School of Electrical, Electronic and Mechanical Engineering, University of Bristol, Bristol BS8 1UB, U.K. (e-mail: shuping.dang@bristol.ac.uk).

Baoxin Li is with the Department of Computer Science, Arizona State University, Tempe, AZ 85287, USA (e-mail: baoxin.li@asu.edu).

knowledge of a specific ISP pipeline, it is challenging to directly learn the operations in the ISP pipeline from a single RAW image. Although one can train the JDSR models using RAW and sRGB image pairs [17], [18], [20], the trained models may exhibit a poor generalization ability.

To address this problem, some RAW image SR methods additionally leverage the compressed LR sRGB images, where the ISP-pipeline-related knowledge is embedded. For instance, Chang *et al.* [13] proposed a two-stage CNN (TSCNN) method, where the LR sRGB image serves as the initial estimation for the first stage of restoration. However, since the compressed LR sRGB image is directly inserted into TSCNN, the accumulated errors may still mislead the following processing procedure. Xu *et al.* [21] proposed to estimate two transformation matrices from the LR sRGB image, and then use them to correct colors. However, it is difficult to accurately model the complex processing steps in the ISP pipeline by using the simple operations like matrix multiplication and addition. Furthermore, in [21], color correction is conducted after SR, and the separation of these two steps could result in less effective recover of image details.

To effectively restore image details and obtain appropriate colors for rendering, we propose a new RAW-image-SR framework called pyramid restoration network (PRNet) in this paper. PRNet contains three sub-networks: an initial reconstruction (IR) sub-network, responsible for estimating an intermediate HR image from the LR RAW input; an auxiliary color guidance (ACG) generator, which produces external guidance using the LR sRGB image; and a pyramid refinement (PR) sub-network, tasked with refining the intermediate result and correcting its colors based on the external guidance. As the PR sub-network is responsible for establishing a sophisticated mapping, the divide-and-conquer strategy is needed to assist the process. To achieve this, we use the Laplacian pyramid decomposition proposed in [22] to divide the intermediate HR result into a low-frequency (LF) component and multiple high-frequency (HF) components. Since color distribution information is mainly depicted by the LF component, we have developed the cross-layer correction module (CLCM) to effectively integrate the external guidance into the LF component for color correction. In CLCM, the multi-head-attention-based modulation (MAM) block explores the long-range interactions between the external guidance and the LF component, and then uses the produced features to accurately compute the scale and shift parameters for affine transformation [23]. To restore sharp edges and rich details, we take advantage of the inter-layer correlation among different frequency components in the Laplacian pyramid. More specifically, the refined features in a lower layer are used to support the refinement of an upper layer. Therefore, the HF components are progressively refined in a coarse-to-fine manner. In comparison to previous approaches, such as [20] and [21], our method conducts color correction and detail enhancement more effectively and efficiently as a unified process.

As shown in Fig. 1, our PRNet significantly outperforms other compared methods while achieving the best trade-off between model complexity and performance. In summary, the main contributions of this paper are four-fold:

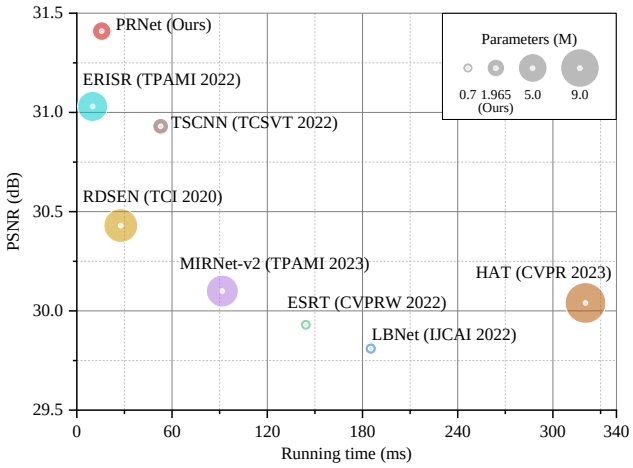


Fig. 1: Performance comparison between PSNR value, model size and running time (×2). For the input with a resolution of 192 × 192, the runtime is measured on a single NVIDIA GeForce RTX 3090 GPU. A bigger circle size stands for a larger number of parameters.

1) We present a new framework named PRNet for RAW images SR, where the PR sub-network applies a pyramid restoration structure to divide the image into various frequency components and restore them in a coarse-to-fine manner. In comparison to previous approaches, our framework simultaneously enhances image details and corrects colors in a more effective and efficient way.
2) To effectively explore the external guidance from LR sRGB images for the reconstruction of the LF component, we design a CLCM and insert it into the LF layer of the PR sub-network. In CLCM, the long-range dependency between two inputs is fully explored by the MAM block, and thus the representation ability of the normal affine transformation is considerably improved.
3) To fully make use of the inter-layer correlation within the Laplacian pyramid, the cross-layer refinement module (CRM) is built by integrating the CLCM. It is able to well facilitate feature refinement in the HF layers with the processed features from a lower layer.
4) Extensive experiments presented in this paper demonstrate that with a relatively low computational cost, the proposed PRNet outperforms other state-of-the-art (SOTA) methods on both synthetic and realistic datasets. To facilitate further study, our source code and the pre-trained models will be released at https://github.com/lingmy0713/PRNet.

The rest of this paper is organized as follows: Section II provides a brief review of the background and related techniques; the proposed PRNet is detailed in Section III; in Section IV, extensive experiments are presented to verify the effectiveness and efficiency of the proposed approach; finally, Section V concludes this paper. For ease of reference, Table I lists the abbreviations used in this paper.

II. Related Background

A. SR on sRGB Images

There are various CNN-based SR approaches which learn the nonlinear mapping from the LR space to the HR space.

TABLE I: List of Abbreviations.

Abbreviations	Meanings
LR/HR/SR	Low-resolution/high-resolution/super-resolution
ISP	Image signal processing
LUT	Look-up table
CFA	Color filter array
JDSR	Joint demosaicking and super-resolution
PRNet	Pyramid restoration network
IR sub-network	Initial reconstruction sub-network
ACG generator	Auxiliary color guidance generator
PR sub-network	Pyramid refinement sub-network
LF component	Low-frequency component
HF component	High-frequency component
RCAB / RCAG	Residual channel attention block/group
MV2	MobileNet-V2 block
SFT	Spatial feature transform
CRM	Cross-layer refinement module
CCB / CLCM	Cross-layer correction block/module
CAB	Cross-layer attention block
MHA	Multi-head attention
MAM	Multi-head-attention-based modulation
GDFN	Gated-dconv feed-forward network

The classical structures and mechanisms include: to build a deep model for sophisticated mapping, residual learning [24] has been widely adopted to mitigate the gradient vanishing problem [25]–[27]; to fully explore multi-layer features and to facilitate information flow, dense connection has been incorporated into several SR models [28]–[30]; to effectively re-scale features, the attention mechanisms have also been applied in many SR models [31]–[34]; in [35], [36], [37] and [38], the multi-scale learning strategy is utilized.

Recently, due to its impressive ability in capturing global interactions between contexts, the *Vision Transformer* developed in [39] has also been successfully introduced to image SR [40]–[47]. However, as computing long-range dependency across the whole image is highly complex, different approaches have been proposed to reduce the computational cost and the required memory. For instance, the shifted window mechanism is applied in [40] and [41] to restrict the computation of attention map in a relatively small region; Lu *et al.* [42] split the *query*, *key*, and *value* into multiple segments; Gao *et al.* [43] developed a lightweight bimodal network (LBNet), where a symmetric CNN is designed for local feature extraction and the recursive transformer is utilized to obtain global information; In [45], cross-covariance is computed across feature channels rather than the spatial dimension.

Although numerous image SR methods have been proposed, the performance of these methods is greatly limited by the sRGB input. This is because sRGB images suffer from low bit depth, have nonlinear relationship with scene radiance, and contain artifacts caused by image compression, demosaicking, and noise reduction. To tackle this problem, we resort to conducting SR on RAW images in this paper.

B. SR on RAW Images

As RAW images only contain one color channel per pixel, color image demosaicking is required in the ISP pipeline to reconstruct the full-color image. To ensure accurate reconstruction, in recent years, many CNN-based demosaicking approaches have been proposed [5]–[9], [49]–[53]. Note that the color channels recorded at vertical or horizontal neighboring positions in RAW images are different. Therefore, before feature extraction, a RAW image is typically rearranged into four quarter-resolution RGGB matrices by packing 2×2 blocks [5]–[7], [51]. Although one can directly cascade an image demosaicking method with an sRGB image SR method, separately conducting these two tasks leads to sub-optimal results, as the artifacts introduced by demosaicking may mislead the following SR method [13].

Since the image demosaicking and SR tasks are highly related, some JDSR methods have been proposed [17]–[20]. The very first CNN-based JDSR method trains a deep residual network in an end-to-end manner [17]. Later on, Xu *et al.* [18] proposed to generate an intermediate demosaicked result by the pre-demosaicking network (PDNet), and then reconstruct the HR full color result by using the residual-dense squeeze-and-excitation networks (RDSEN). Xing *et al.* [19] developed a joint image denoising, demosaicking and SR network (JDnDmSR), where multiple residual channel attention blocks (RCABs) are cascaded [32]. Qian *et al.* [20] proposed a trinity pixel enhancement network (TENet), which adopts residual in residual dense blocks (RRDBs) developed in [26] to boost performance. To train and evaluate RAW image SR methods, a realistic dataset called *PixelShift200* was established in [20] by using the pixel shift technology. Zhang *et al.* [54] contributed a dataset named *SR-RAW*, where the ground-truth (GT) data is obtained via optical zoom. However, since there exists an obvious misalignment between the image contents captured by different focal lengths, a contextual bilateral loss is required when training SR models on the *SR-RAW* dataset [54]. Although the above techniques partially solve the error accumulation problem, it is still difficult to directly learn a mapping relation from the LR linear space to the HR sRGB space, as RAW images do not contain prior knowledge of the specific ISP pipeline. As a result, the learned mapping relation cannot be well generalized from one camera style to another.

Due to the fact that digital cameras also output the sRGB images pre-processed by their own ISP pipelines, Xu *et al.* [21], [55] built a dataset containing two types of LR images, i.e., the RAW images and the corresponding JPEG format sRGB images (for simplicity, this dataset is called *RAW-sRGB* in this paper). Furthermore, a two-branch CNN structure called ERISR (exploiting RAW images for SR) is developed in [21], [55], where one branch recovers fine details from the RAW input while the other estimates two matrices from the LR sRGB image for color correction. Chang *et al.* [13] proposed a two-stage CNN architecture (TSCNN), where the first stage of network reconstructs an initial result of full color image and then the second stage further enhances its spatial resolution and suppresses noises and artifacts. When training TSCNN on *RAW-sRGB*, the compressed sRGB image is inserted into the first stage of the network to facilitate residual learning.

In this paper, we also present a RAW image SR method that utilizes both the RAW image and the compressed sRGB image

Fig. 2: sRGB images with different camera styles. From left to right: the RAW image, the sRGB images produced by FUJIFILM GFX100S, Nikon D700, Canon 5D Mark II, Sony ILCE-7RM4 and Leica M8. These images are generated by applying Adobe DNG software development kit with different metadata on the RAW image from the *MIT-Adobe FiveK* dataset [48].

as input. Nevertheless, in our PRNet, the initially reconstructed image is decomposed to different frequency components, and each component is enhanced either under the external guidance from the sRGB input or under the inter-layer guidance from a lower pyramid layer. Our structure results in more effective and efficient color correction and detail enhancement.

III. Proposed Method

A. Framework Design

Besides breaking through the resolution limitations of camera sensors, conducting SR on RAW images also requires establishing an accurate mapping relation from the linear color space to the nonlinear sRGB space. However, the RAW images do not contain the specific information of the ISP pipeline. In addition, the system and parametric designs are typical withheld by camera manufactures, making the real ISP pipeline practically a "black box". Therefore, it is difficult to apply a CNN structure to directly learn such a complex mapping function.

On the other hand, for different types and brands of cameras, the processing steps in the ISP pipeline, such as LUT and color space transformation, are distinct from each other. As a result, a RAW image can be rendered to multiple styles of sRGB images, making the RAW-to-sRGB transformation a one-to-many mapping. For better understanding, Fig. 2 shows some examples of different camera styles. Consequently, without knowing the precise camera style information before hand, it is difficult for the CNN structure to generalize from one camera style to another.

Fortunately, besides RAW images, digital cameras provide sRGB images to users, e.g., stored in JPEG format. Since these sRGB images contain specific information of the ISP pipeline, it is reasonable to learn color features from them and then use these learned features to assist in building a sophisticated mapping function from the LR RAW space to the HR sRGB space. Following this thought, we design the PRNet as shown in Fig. 3, which illustrates the three key components: an initial reconstruction (IR) sub-network, a pyramid refinement (PR) sub-network, and an auxiliary color guidance (ACG) generator.

Firstly, given an LR RAW input $\mathbf{X}_{\text{raw}} \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times 1}$, the IR sub-network generates a rough estimation of the HR full color image $\mathbf{X}_{\text{lin}} \in \mathbb{R}^{H\times W\times 3}$ in the linear color space, where H and W denote the height and width of the HR image, respectively, and s is the scale factor. Fig. 3 shows the detailed structure of IR sub-network. Similar to [5]–[7], the RAW image $\mathbf{X}_{\text{raw}}$ is rearranged to a quarter-resolution image $\mathbf{X}_{\text{pack}} \in \mathbb{R}^{\frac{H}{2s}\times\frac{W}{2s}\times 4}$ according to the Bayer pattern of RGGB. To further refine deep features, two residual channel attention groups (RCAG) are cascaded [32]. To increase the spatial resolution, a sub-pixel layer can be placed at the end of the IR sub-network [56].

Next, the PR sub-network further enhances the details of $\mathbf{X}_{\text{lin}}$ and builds a sophisticated mapping from the linear color space to the nonlinear color space, so as to obtain the final HR sRGB result, i.e., $\mathbf{X}_{\text{srgb}} \in \mathbb{R}^{H\times W\times 3}$. The PR sub-network is the most important part in the PRNet, the structure of which is detailed in Sections III-B, III-C, and III-D.

To accurately correct colors, the ACG generator extracts the ISP-pipeline-related color features, denoted as $\mathbf{G}$, from the compressed LR sRGB image, i.e., $\mathbf{X}_{\text{ref}} \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times 3}$. Then, $\mathbf{G}$ is fed to the PR sub-network as external guidance. Similar to the IR sub-network, RCAG is also applied in the ACG generator to refine features [32]. Nevertheless, a MobileNet-V2 (MV2) block is employed in the ACG generator to reduce the spatial resolution of features [57], followed by a sigmoid layer to keep the output within the range of 0 to 1.

Note that both the IR sub-network and the ACG generator only contain RCAGs, rather than other computationally intense models, such as transformer [58]. The main reasons are: 1) as roughly estimated HR image $\mathbf{X}_{\text{lin}}$ will be further refined by the following PR sub-network, it is more efficient to allocate more resources to the PR sub-network; 2) further introducing the transformer mechanism to the IR sub-network or the ACG generator could significantly increase the computational burden of PRNet.

In summary, the processing steps of the proposed framework can be represented as

$$\mathbf{X}_{\text{srgb}} = \mathcal{N}_{\text{PR}}(\mathcal{N}_{\text{IR}}(\mathbf{X}_{\text{raw}}) \mid \mathbf{G}), \tag{1}$$

where $\mathcal{N}_{\text{IR}}(\cdot)$ and $\mathcal{N}_{\text{PR}}(\cdot)$ are the functions of the IR and PR sub-networks, respectively; $\mathbf{G}$ serves as prior knowledge and is obtained by

$$\mathbf{G} = \mathcal{F}_{\text{ACG}}(\mathbf{X}_{\text{ref}}), \tag{2}$$

where $\mathcal{F}_{\text{ACG}}(\cdot)$ stands for the function of the ACG generator.

To train the PRNet, the following loss function is used:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{ssim} + \mathcal{L}_{pyr}, \tag{3}$$

where $\mathcal{L}_{rec}$, $\mathcal{L}_{ssim}$ and $\mathcal{L}_{pyr}$ represent the reconstruction loss, the SSIM loss and the pyramid loss, respectively. The pyramid loss will be defined in Section III-B. The reconstruction loss is computed by

$$\mathcal{L}_{rec} = \|\mathbf{X}_{\text{srgb}} - \mathbf{X}_{\text{gt}}\|_1, \tag{4}$$

where $\mathbf{X}_{\text{gt}}$ and $\mathbf{X}_{\text{srgb}}$ denote the GT HR sRGB image and the image reconstructed by PRNet, respectively. The SSIM loss can be defined as

$$\mathcal{L}_{ssim} = 1 - SSIM(\mathbf{X}_{\text{srgb}}, \mathbf{X}_{\text{gt}}), \tag{5}$$

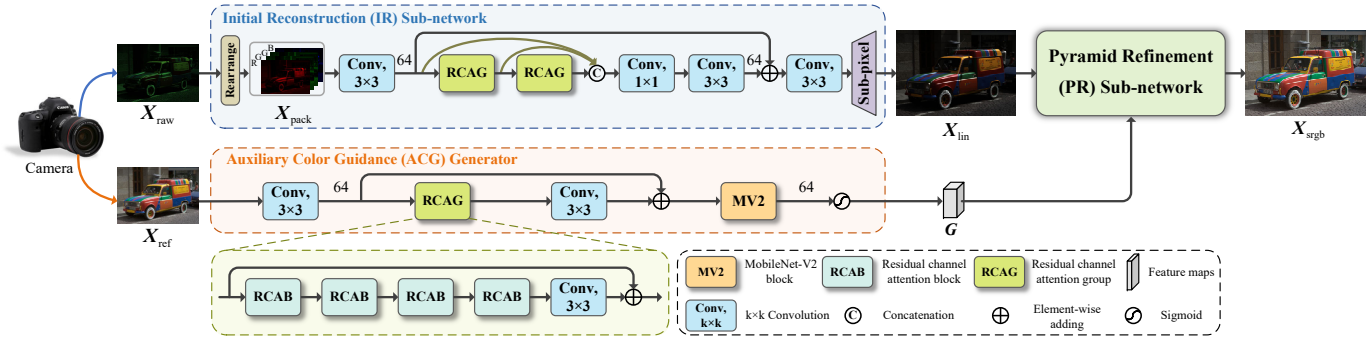


Fig. 3: Overall structure of the proposed PRNet. The numbers on lines indicate the numbers of feature channels.

where function $SSIM(\cdot)$ computes the SSIM value of the reconstructed result.

It should also be pointed out that similar to TSCNN [13], we do not apply GT HR linear images to supervise the estimation of $\mathbf{X}_{\text{lin}}$ during training. The reasons are: 1) it is not necessary to accurately estimate $\mathbf{X}_{\text{lin}}$, as details and colors of an image will be further restored/corrected by the following PR sub-network; 2) additionally applying GT HR linear images to constraint the estimation of $\mathbf{X}_{\text{lin}}$ could diminish the importance of the final reconstruction loss (c.f., Eq. (3)), which may lead to less accurate results.

B. Laplacian-Pyramid-Based Refinement

To ease the difficulty of simultaneously enhancing details and correcting colors of $\mathbf{X}_{\text{lin}}$, we apply the divide-and-conquer strategy in the PR sub-network. More specifically, as shown in Fig. 4, the Laplacian pyramid decomposition [22] is introduced to decompose $\mathbf{X}_{\text{lin}}$ into different components, and the components with different spatial resolutions are refined in a coarse-to-fine manner.

To better illustrate the effect of Laplacian pyramid decomposition, the decomposed results of a pair of linear and sRGB images are shown in Fig. 5. It can be seen from this figure that: 1) the colors are mainly represented by the LF component in the pyramid, which encourages us to focus on correcting colors in the lowest pyramid layer; 2) the HF components with higher resolutions capture richer textures and sharper edges, which suggests that it is crucial to enhance the HF components for the resolution-sensitive tasks; 3) the co-located positions in various frequency components with different scales are highly correlated, which inspires us to exploit inter-layer correlation among different pyramid layers.

Based on the above analysis, the Laplacian pyramid decomposition is carried out on $\mathbf{X}_{\text{lin}}$, leading to

$$[\mathbf{H}_1, \mathbf{H}_2, \cdots, \mathbf{H}_n, \mathbf{L}] = \mathcal{D}(\mathbf{X}_{\text{lin}}), \tag{6}$$

where $\mathcal{D}(\cdot)$ denotes the function of Laplacian decomposition; $\mathbf{H}_i \in \mathbb{R}^{\frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}} \times 3} (1 \le i \le n)$ and $\mathbf{L} \in \mathbb{R}^{\frac{H}{2^n} \times \frac{W}{2^n} \times 3}$ are the i-th HF component and the LF component, respectively; n is the number of HF layers in the Laplacian pyramid. Note that the Laplacian pyramid decomposition is fully revertible, and only requires simple blurring and downsampling operations with very limited computational cost.

As discussed before, the colors are mainly represented by the LF component. Therefore, the CLCM is developed and incorporated into the LF layer, so that the LF component can be properly corrected under the external guidance from $\mathbf{G}$. On the other hand, precisely enhancing the HF components is crucial for restoring small structures and fine details in images. Since there exists a strong spatial correlation among different frequency components, to facilitate the reconstruction of HF components, exploring inter-layer correlation in Laplacian pyramid is necessary. Thus, CRM is applied in each HF layer, which leverages the component from a lower layer to support the refinement in the current processing layer. The correction of the LF component and the refinement of the HF components are detailed in Section III-C and Section III-D, respectively.

Finally, the super-resolved sRGB result $\mathbf{X}_{\text{srgb}}$ can be obtained by applying Laplacian pyramid reconstruction on the refined components, i.e.,

$$\mathbf{X}_{\text{srgb}} = \mathcal{R}([\hat{\mathbf{H}}_1, \hat{\mathbf{H}}_2, \cdots, \hat{\mathbf{H}}_n, \hat{\mathbf{L}}]), \tag{7}$$

where $\mathcal{R}(\cdot)$ represents the function of Laplacian pyramid reconstruction; $\{\hat{\mathbf{H}}_i\}$ and $\hat{\mathbf{L}}$ denote the refined HF and LF components, respectively.

Inspired by [59], to provide a more reliable reconstruction, we also introduce the pyramid loss to constraint the reconstruction of all pyramid components during training. However, the architecture of our PR sub-network greatly differs from the model in [59]. Therefore, unlike [59], which measures the error of a restored image at each pyramid layer, for the HF layers, we directly impose the minimization of the reconstruction error of residual images (i.e., $\{\hat{\mathbf{H}}_i\}$ in Eq. (7)). By doing so, the local sharpness of HF components can be guaranteed. Thus, the pyramid loss in Eq. (3) can be determined as

$$\mathcal{L}_{pyr} = \sum_{i=1}^{n} \|\hat{\mathbf{H}}_i - \mathbf{H}_i^*\|_1 + \|\hat{\mathbf{L}} - \mathbf{L}^*\|_1, \tag{8}$$

where $\{\mathbf{H}_i^*\}$ and $\mathbf{L}^*$ stand for the HF and LF components that are decomposed from the GT HR sRGB image by using Eq. (6), respectively.

C. Correction of the LF Component in Laplacian Pyramid

As mentioned before, the RAW data only records the radiance information captured by sensors. To enhance the generalization ability of our model, effectively utilizing the external guidance is necessary for proper color correction. Therefore, feature $\mathbf{G}$, which is extracted from the LR sRGB image pre-processed by the ISP pipeline, is fed to the CLCM

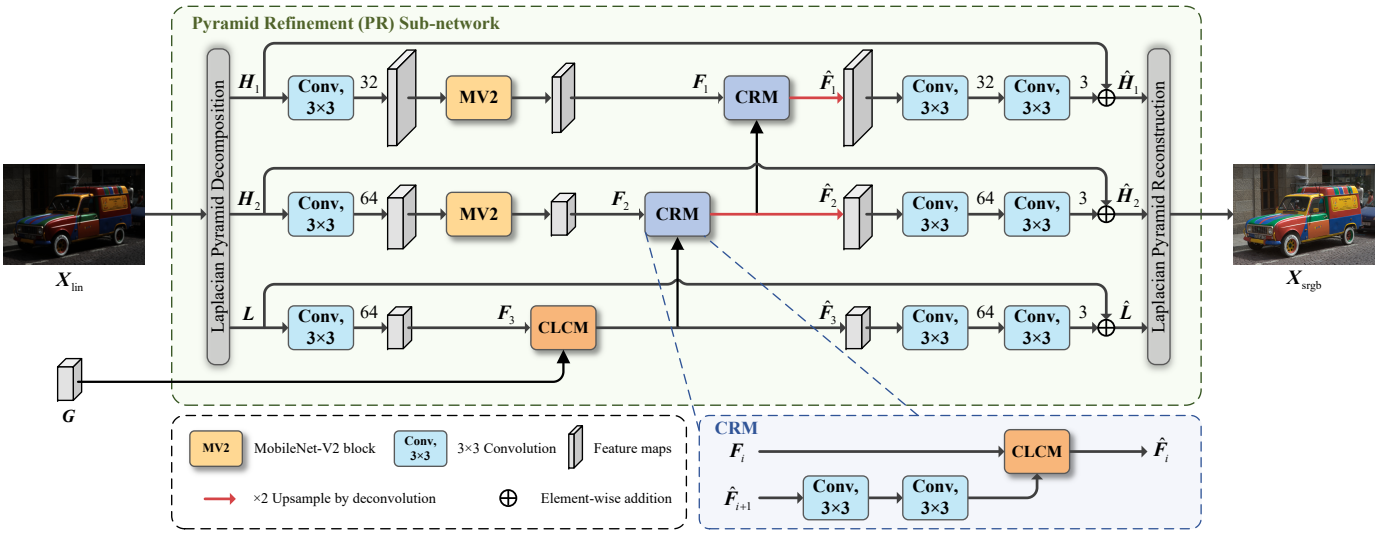
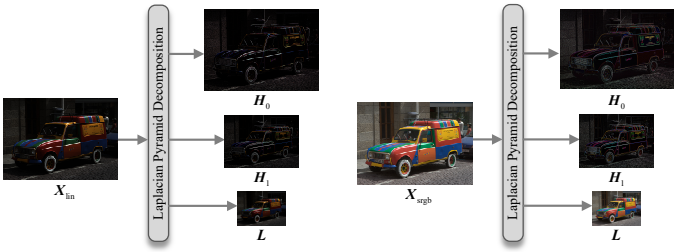


Fig. 4: Structure of the PR sub-network. Here, we show three layers of Laplacian pyramid decomposition as an example, and the numbers on lines indicate the numbers of feature channels.



(a) Image in the linear color space. (b) Image in the sRGB space.

Fig. 5: Laplacian pyramid decomposition of a pair of linear and sRGB images. Here we only show three layers of decomposition: $\mathbf{H}_1$, $\mathbf{H}_2$ are the first and second layers of HF components, and $\mathbf{L}$ is the LF component. The image in the linear color space, i.e., $\mathbf{X}_{\text{lin}}$, is provided in the *RAW-sRGB* dataset [21].

for correcting the LF component in the pyramid. Fig. 6 gives the detailed structure of CLCM, where m cross-layer correction blocks (CCBs) are cascaded to progressively correct LF feature $\mathbf{F}$ under the external guidance from $\mathbf{G}$. To fully exploit the hierarchical features, the features fed to each CCB are also densely connected to the output of the last CCB, followed by a 1×1 convolution layer to aggregate and compress the concatenated features.

To effectively and efficiently incorporate the external guidance, affine transformation used in [23] may be considered as a possible solution. By applying affine transformation, the current feature $\mathbf{F} \in \mathbb{R}^{h \times w \times c}$, with h, w and c denoting the height, width and number of channels of the processing features, can be modulated to a transformed feature $\mathbf{Y}$ as

$$\mathbf{Y} = \boldsymbol{\gamma} \odot \mathbf{F} + \boldsymbol{\beta}, \tag{9}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are the scaling and shifting parameters, respectively, each of which shares the same dimension as $\mathbf{F}$ and can be learned from external guidance $\mathbf{G}$; $\odot$ denotes element-wise multiplication.

Note that in traditional approaches, scaling parameter $\boldsymbol{\gamma}$ and shift parameter $\boldsymbol{\beta}$ are usually learned by applying normal convolutions on the external features [23]. However, normal convolutions have very limited receptive fields, which considerably hinders the network from capturing long-range dependence of image contents. Consequently, the effectiveness of affine transformation could be restricted by the inappropriate scaling and shifting parameters.

To address this issue, the *Transformer* mechanism is introduced, leading to the structure of CCB shown in Fig. 6. As can be seen, CCB contains an MAM block, which aims to module feature $\mathbf{F}$ by fully investigating the long-range interactions between $\mathbf{F}$ and $\mathbf{G}$. To facilitate the useful information flow passing through the network, the gated-dconv feed-forward network (GDFN) proposed in [45] is directly applied after the MAM block. In the front of the MAM block, we generate *query* ($\mathbf{Q}$) from the layer normalized feature $\mathbf{F}$, and meanwhile produce *key* ($\mathbf{K}$) and two *values* ($\mathbf{V}_\beta$ and $\mathbf{V}_\gamma$) from layer normalized feature $\mathbf{G}$, respectively. These projections are obtained by applying 1×1 convolutions to aggregate channel features, followed by 3×3 depth-wise convolutions to enrich local information.

To form a multi-head attention mechanism in the MAM block, we reshape $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}_\gamma$, and $\mathbf{V}_\beta$ and divide them into k "heads" [1]. Let us take $\mathbf{Q}$ as an example to illustrate. Supposing that $\mathbf{Q} \in \mathbb{R}^{h \times w \times c}$, $\mathbf{Q}$ is reshaped to the form of $[\hat{\mathbf{Q}}_1, \cdots, \hat{\mathbf{Q}}_k]$, with $\hat{\mathbf{Q}}_i \in \mathbb{R}^{hw \times \frac{c}{k}}$ denoting the i-th head of $\mathbf{Q}$. For each head, a cross-attention map is learned separately. However, considering the heavy computational burden of calculating attention across spatial dimension, we resort to computing cross-covariance over feature channels. As a result, the attention map for the i-th head can be formulated as:

$$\mathbf{M}_i = \mathcal{S}(\hat{\mathbf{K}}_i \cdot \hat{\mathbf{Q}}_i / \alpha_i), \tag{10}$$

where $\mathcal{S}(\cdot)$ represents the softmax function to generate attention scores; α_i is a learnable temperature parameter that adaptively scales the matrix multiplication.

Afterwards, the attention map, denoted as $\mathbf{M}_i \in \mathbb{R}^{\frac{c}{k} \times \frac{c}{k}}$, is multiplied with the i-th head of $\mathbf{V}_\gamma$ and $\mathbf{V}_\beta$, i.e., $\hat{\mathbf{V}}_{\gamma i}$ and $\hat{\mathbf{V}}_{\beta i}$, thus yielding two output features $\hat{\mathbf{A}}_{\gamma i}$ and $\hat{\mathbf{A}}_{\beta i}$, respectively. By concatenating the output features from all k heads, we obtain $\mathbf{A}_\gamma = [\hat{\mathbf{A}}_{\gamma 1}, \cdots, \hat{\mathbf{A}}_{\gamma k}]$ and $\mathbf{A}_\beta = [\hat{\mathbf{A}}_{\beta 1}, \cdots, \hat{\mathbf{A}}_{\beta k}]$,

[1] k is set as 4 in this paper, which is the same as [45].

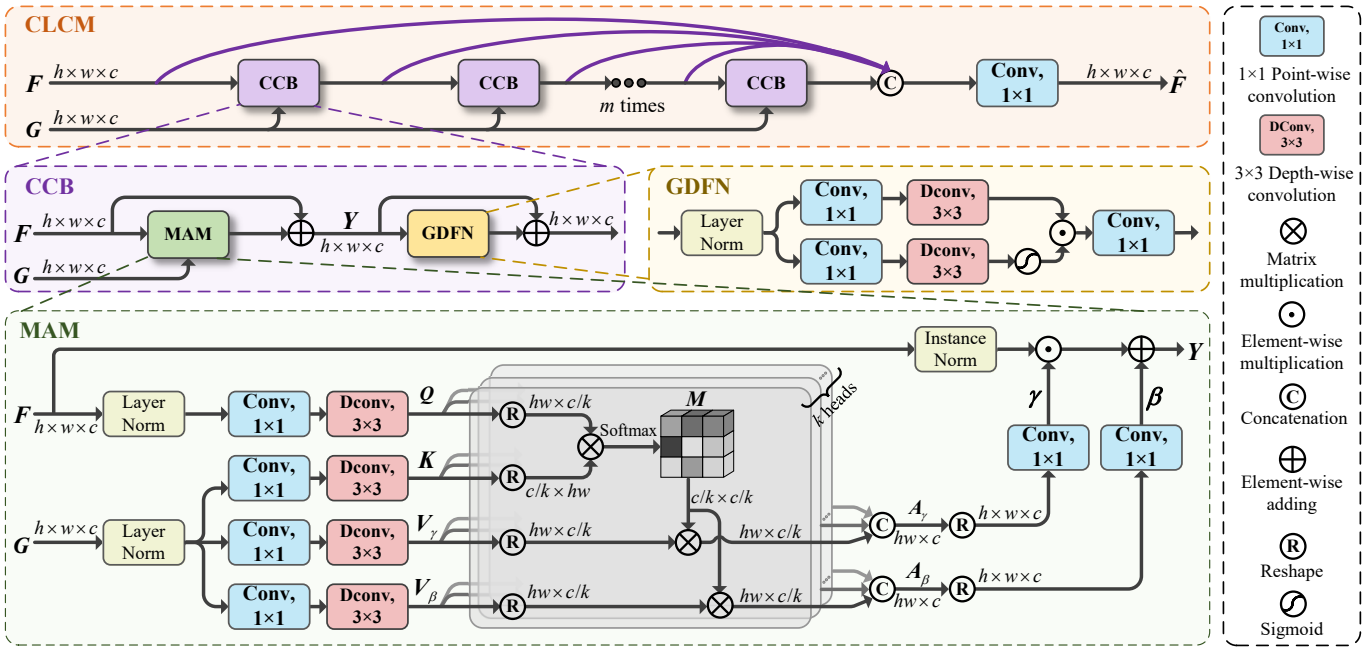


Fig. 6: Structures of CLCM and CCB.

with $\mathbf{A}_\gamma$ and $\mathbf{A}_\beta \in \mathbb{R}^{hw\times c}$. Finally, scaling and shifting parameters $\boldsymbol{\gamma} \in \mathbb{R}^{h\times w\times c}$ and $\boldsymbol{\beta} \in \mathbb{R}^{h\times w\times c}$ are obtained by

$$\boldsymbol{\gamma} = \mathcal{C}_1\left(\mathcal{F}_{\mathrm{R}}\left(\mathbf{A}_\gamma\right)\right), \tag{11}$$

and

$$\boldsymbol{\beta} = \mathcal{C}_1(\mathcal{F}_{\mathrm{R}}(\mathbf{A}_\beta)), \tag{12}$$

where $\mathcal{C}_1$ denotes a 1×1 convolution layer, and function $\mathcal{F}_{\mathrm{R}}(\cdot)$ reshapes an $hw\times c$ dimensional matrix to an $h\times w\times c$ tensor.

With the parameters obtained by Eqs. (11) and (12), the features in the LF pyramid layer can be modulated by Eq. (9). Note that to avoid significant fluctuations in statistical values of each instance, the instance normalization developed in [60] is adopted before applying the affine transformation.

It should be pointed out that although ERISR reported in [21] also carries out color correction under external guidance, our approach differs significantly from it, and the main differences are summarized below: 1) ERISR directly performs color correction on each pixel, while the colors are mainly corrected in the LF layer of the Laplacian pyramid in our PR sub-network. Such a divide-and-conquer strategy enables our PR sub-network to have a stronger representation ability, thus leading to more reliable color correction. 2) Only 3×3 matrix multiplications and additions are used to perform color correction in ERISR. However, it is difficult for these simple operations to precisely imitate the complex processing steps, such as LUT and tone curves in the ISP pipeline. On the contrary, the color correction is progressively processed by m cascaded CCBs in the CLCM of the PR sub-network, which is a more sophisticated process. 3) In ERISR, color correction is applied after the image restoration. In contrast, color correction and detail enhancement are jointly conducted in the PR sub-network. Due to the effective interaction between these two sub-tasks in the PR sub-network, the problem of error accumulation can be significantly reduced, and the refinement of HF components can also greatly benefit from the accurate color correction.

D. Refinement of the HF Components in Laplacian Pyramid

To fully explore inter-layer correlation within the Laplacian pyramid, we design the coarse-to-fine reconstruction structure shown in Fig. 4. To well capture long-range correlations between adjacent pyramid components, a CRM is applied in each HF layer. The CRM contains a CLCM to perform guided feature refinement, where the refined features from the lower layer are used as the guidance for the current layer. To prepare an appropriate guidance, two 3×3 convolution layers are employed to adjust the refined features from the lower layer before feeding them to the CLCM.

As a higher pyramid layer leads to a larger spatial resolution, the computational cost increases rapidly in the high pyramid layers, especially for the MAM block where the *Transformer* mechanism is introduced. To mitigate the computational burden, the MV2 block designed in [57] is adopted in each HF layer, which downsamples the shallow features by a factor of 2. Compared to other downsampling operations, the main advantage of MV2 block lies in that it is able to preserve more important features with low computational cost. More specifically, in an MV2 block, a 1×1 point-wise convolution layer expands the number of feature channels, and then a 3×3 depth-wise convolution layer with a stride of 2 is used to downsample features, followed by another 1×1 point-wise convolution layer to compress the expanded feature channels back to the original dimension. Due to the superior performance in downsampling, the MV2 block has been widely used in other works, such as [61], [62]. Nevertheless, in resolution-sensitive tasks, largely reducing the spatial resolution of features could be harmful for restoring fine details. To achieve a better trade-off between complexity and representation ability, only one MV2 block is applied before the CRM, and then the feature

refined by CRM is upsampled by a deconvolution layer to restore its original spatial resolution.

In the PR sub-network, the number of pyramid layers is set as 3, as we observe that more pyramid layers do not necessarily lead to higher performance (the experimental results can be found in Section IV-C). From the bottom layer to the top one, the feature channels are 64, 64, and 32, respectively, and the numbers of CCBs in a CLCM are set as 4, 2, and 2, respectively. Compared to the lower pyramid layers, we allocate less parameters for the upper ones because: 1) with the guidance from the well refined features in lower layers, the difficulty of feature refinement in upper layers has been largely reduced; 2) the high spatial resolution of the components in upper layers could result in a heavy computational burden if allocated with parameters more than necessary.

IV. Experimental Results

A. Experimental Settings

The synthetic dataset *RAW-sRGB* reported in [21] is used for evaluation in this paper. This dataset is synthesized from the well-known dataset *MIT-Adobe FiveK* [48], consisting of 1300 LR/HR image pairs for training and 150 LR/HR image pairs for testing. Each image pair includes: 1) 10 LR RAW images that are first blurred by different defocus or motion kernels, and then contaminated by heteroscedastic Gaussian noise and dowsampled by the factors of 2 and 4; 2) 10 compressed LR sRGB images generated from the LR RAW images (in JPEG format); 3) one GT HR sRGB image.

In addition, we also evaluate different methods on a realistic dataset [21], where 100 images ranging from 3024×4032 to 6208×8736 are captured by 7 brands of cameras, including Nikon, Canon, Sony, Leica, Pentax, Fuji, and Apple (iPad Pro and iPhone 6s Plus). The Rawpy toolkit (a Python version of LibRaw) is used to produce the LR JPEG image. All images in the realistic dataset are used for testing.

The full-reference quality metrics, including color peak-signal-to-noise ratio (CPSNR) [63], structure similarity (SSIM) index [64], and learned perceptual image patch similarity metric (LPIPS) [65], are used to measure the performance of competing methods on the synthetic dataset, while the no-reference quality metrics, including natural image quality evaluator (NIQE) [66], perception-based image quality evaluator (PIQE) [67], no-reference quality metric (NRQM) [68], and perceptual index (PI) [69], are applied when testing on the realistic dataset. Note that higher values of PSNR, SSIM, NRQM, and PI, and lower values of LPIPS, NIQE, and PIQE indicate better image quality.

When training on synthetic dataset, following [21], we randomly crop the RAW images and the compressed LR sRGB images into 256×256 patches at the scale factor of $\times 2$, and 128×128 patches at the scale factor of $\times 4$. No data augmentation is used. The batch size is set to be 6, and the AdamW optimizer [70] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay $\omega = 0.05$ is used to train our PRNet for 300 epochs. The warmup strategy [24] is applied, and the learning rate is linearly increased to 1.5×10^{-4} and then dropped to 6×10^{-6} by using the cosine annealing strategy.

All experiments are conducted with the PyTorch framework on a single NVIDIA GeForce RTX 3090 GPU. To reduce the required GPU memory, when testing on images at the scale factors of $\times 2$ and $\times 4$, the input image is cropped to overlapped 256×256 and 128×128 patches, respectively. After being processed by the PRNet, the patches are merged back to a super-resolved sRGB image.

B. Comparison with SOTA Methods

To demonstrate the effectiveness of the proposed PRNet, three types of methods are compared as benchmarks in this paper, which are: 1) the JDSR methods which directly learn a nonlinear mapping from the LR linear RAW space to the HR sRGB full color space (denoted as *JDSR w/o sRGB input*), including JDnDmSR [19] and TENet [20]; 2) the SR methods which are applied on the LR sRGB images pre-processed by the in-camera ISP pipeline (denoted as *sRGB image SR*), including LBNet [43], ESRT [42], HAT [47], and MIRNet-v2 [38]; 3) the RAW image SR methods which jointly make use of the LR RAW images and the LR sRGB images (denoted as *RAW image SR*), including RDSEN [18], ERISR [21], and TSCNN [13]. For a fair comparison, all the compared methods have been re-trained on the *RAW-sRGB* dataset [21]. Note that in TSCNN and RDSEN, the initial interpolation result is replaced by the pre-processed LR sRGB image, and a 6-M parameter version of RDSEN is re-trained in order to obtain a model size similar to other *RAW image SR* methods.

The quantitative comparisons of different methods on the *RAW-sRGB* dataset [21] for the scale factors of $\times 2$ and $\times 4$ are provided in Table II. The following key observations can be summarized from the quantitative results given in this table:

1) The *JDSR-w/o-sRGB-input* methods are significantly inferior to the other two types of methods. This is mainly because the RAW images do not contain any ISP-pipeline-related information, making simultaneously correcting colors and enhancing resolution a challenging task.
2) As SR is performed on the LR sRGB images, no color correction is required for the *sRGB-image-SR* methods. As a result, compared with the *JDSR-w/o-sRGB-input* methods, obviously better performance can be achieved by LBNet, ESRT, HAT, and MIRNet-v2. However, the LR sRGB images are compressed, and only 8 bits are used to record a color channel at each pixel, largely limiting the quality of the super-resolved results.
3) RDSEN, TSCNN, ERISR, and our proposed PRNet are superior to the *sRGB-image-SR* methods, which demonstrates that fully exploring both the RAW image and the LR sRGB image is necessary.
4) Our PRNet achieves the best performance in terms of all objective metrics. Specifically, for the scale factor of $\times 2$, it outperforms the second best method, i.e., ERISR, by 0.38 dB in PSNR, 0.0131 in SSIM, and 0.0190 in LPIPS, respectively. The superior performance of PRNet is attributed to the strong representation ability of the divide-and-conquer strategy, the accurate correction of the LF pyramid component under the external guidance,

TABLE II: Comparison of different methods on the *RAW-sRGB* dataset [21]. The values in bold indicate the best results.

Method	Category	$\times 2$			$\times 4$		
		PSNR (dB) ↑	SSIM ↑	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
JDnDmSR [19]	JDSR w/o sRGB input	21.53	0.7414	0.4382	21.37	0.7209	0.5351
TENet [20]	JDSR w/o sRGB input	21.35	0.7439	0.3898	21.32	0.7219	0.4693
LBNet [43]	sRGB image SR	29.81	0.7780	0.4153	28.56	0.7468	0.4887
ESRT [42]	sRGB image SR	29.93	0.7803	0.4094	28.62	0.7479	0.4838
HAT [47]	sRGB image SR	30.04	0.7828	0.3977	28.66	0.7486	0.4783
MIRNet-v2 [38]	sRGB image SR	30.10	0.7837	0.3962	28.77	0.7510	0.4689
RDSEN [18]	RAW image SR	30.43	0.7951	0.3682	29.59	0.7711	0.4267
TSCNN [13]	RAW image SR	30.94	0.8059	0.3440	29.81	0.7747	0.4090
ERISR [21]	RAW image SR	31.03	0.8083	0.3384	29.75	0.7739	0.4067
PRNet (ours)	RAW image SR	**31.41**	**0.8214**	**0.3194**	**30.11**	**0.7854**	**0.3932**

and the precise refinement of the HF pyramid components by utilizing inter-layer correlation.

Fig. 7 and Fig. 8 show examples of visual results for the scale factors of $\times 2$ and $\times 4$, respectively. It can be observed from both figures that: 1) obvious color distortions can be found in the results yielded by the two *JDSR-w/o-sRGB-input* methods, which demonstrates the prior knowledge from the pre-processed LR sRGB image is essential for faithful color correction; 2) although LBNet, ESRT, HAT, and MIRNet-v2 are able to produce correct colors, they tend to deliver blurred results due to the insufficient information contained in the compressed LR sRGB image; 3) compared to the *sRGB-image-SR* methods, RDSEN, TSCNN, and ERISR have better quality of results, but still cannot produce sharp edges and rich details; 4) our PRNet has the most visually appealing images, of which colors are vivid and details are clear.

C. Ablation Study and Discussions

In this subsection, we conduct ablation study to demonstrate the effectiveness of the proposed method. All experiments are performed on the *RAW-sRGB* dataset at a scale factor of $\times 2$.

Effectiveness of the Modules in PR Sub-Network: The contributions of Laplacian pyramid decomposition, CLCM, and CRM are verified by the quantitative results presented in Table III. The baseline model is built by replacing the whole PR sub-network with cascaded RCAGs [32]. As CLCM is not included in the baseline model, external guidance $\mathbf{G}$ is also removed. For the variants without CLCM or CRM, these modules are also replaced by cascaded RCAGs. To achieve a fair comparison, all the variants are adjusted to a similar model size. As can be seen: 1) variant $\mathbb{N}_a$ has significant improvement over the baseline model, which demonstrates the effectiveness of the Laplacian pyramid decomposition in restoring details and small structures; 2) compared to variant $\mathbb{N}_b$ and the full model, the baseline model and variant $\mathbb{N}_a$ both show dramatic degradation in performance; 3) variant $\mathbb{N}_b$ obtains significant improvement over variant $\mathbb{N}_a$, which implies that CLCM is able to effectively incorporate the external guidance from the LR sRGB images; 4) our full model outperforms variant $\mathbb{N}_b$, which well demonstrates the necessity of exploring inter-layer correlation among the decomposed pyramid layers. From the visual results in Fig. 9, it is obvious that both the baseline and variant $\mathbb{N}_a$ produce false colors due to the removal of external guidance. The full model not only obtains accurate colors, but also preserves more small structures and details than the other variants.

Comparison with the ISP Pipeline Simulator: To further evaluate the effectiveness of our framework, PRNet is compared with a strategy which applies an ISP pipeline simulator, rather than the ACG generator and the PR sub-network, on the output of the IR sub-network. To obtain a precise HR image in the linear color space, the IR sub-network in this alternative strategy has been re-trained under the supervision of GT HR linear images. Note that, although the color correction matrix in the ISP pipeline can be obtained from the metadata in DNG files, the crucial parameters in the processing steps of LUT and tone curves are held by camera manufactures. Therefore, it is impossible to precisely simulate these processing steps in the ISP pipeline. As an alternative, we use the LUT learning strategy proposed by Zeng *et al.* [71], and re-train the model on GT linear-sRGB image pairs in the RAW-sRGB dataset [21]. The quantitative results are listed in Table IV, and the visual comparison is provided in Fig. 10. We can clearly see that the strategy of applying an ISP simulator on the output of the IR sub-network is obviously inferior to our approach. Moreover, the difficulty of achieving a precise ISP simulator also justifies the superiority of our PRNet.

Structure of JPEG-guided RAW Image SR Reconstruction: To validate the effectiveness of the structure of JPEG-guided RAW image SR reconstruction, two variants named as $\mathbb{N}_c$ and $\mathbb{N}_d$, are compared. Variant $\mathbb{N}_c$ is obtained by feeding the LR JPEG image, rather than the LR RAW image, to the IR sub-network. Since there is no need to correct colors in LR JPEG images, the ACG generator is removed from variant $\mathbb{N}_c$. As a result, variant $\mathbb{N}_c$ becomes an sRGB image SR method. For a fair comparison, the model size of $\mathbb{N}_c$ has been adjusted to be similar to that of PRNet. In variant $\mathbb{N}_d$, the inputs of the IR sub-network and the ACG generator are exchanged. Therefore, in contrast to the structure of JPEG-guided RAW image SR reconstruction, $\mathbb{N}_d$ uses the RAW image to guide the SR reconstruction of the JPEG input. From the quantitative results shown in Table IV, it can be found that without the ACG generator, $\mathbb{N}_c$ suffers from a notable

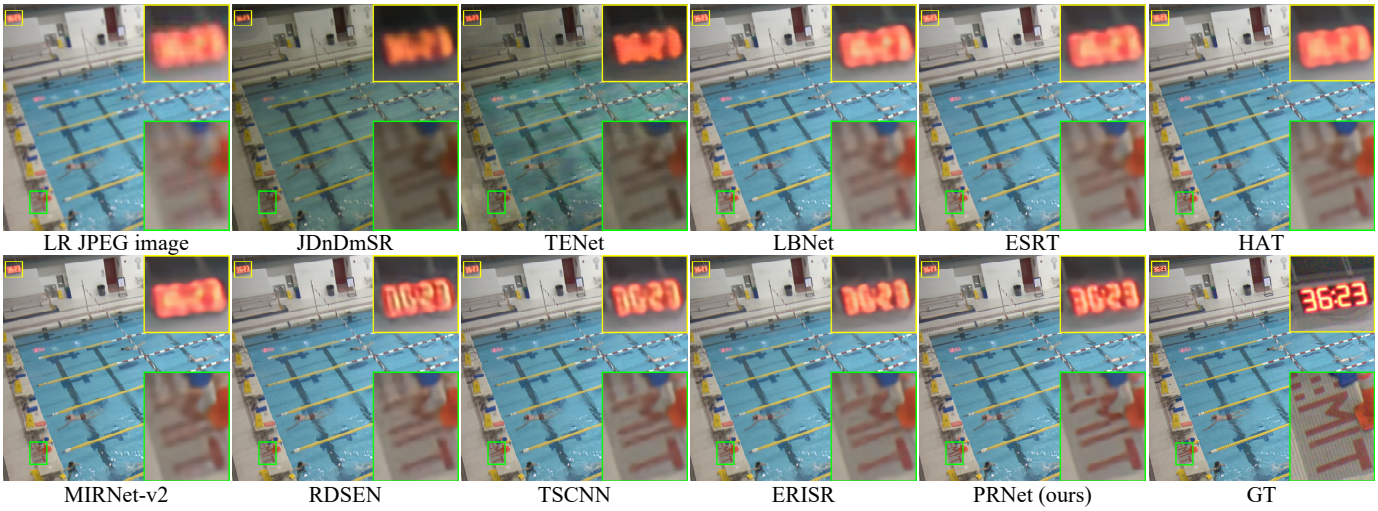


Fig. 7: Qualitative comparison of different methods on image *a0022-IMG_23800000* from the *RAW-sRGB* dataset ($\times 2$).

Fig. 8: Qualitative comparison of different methods on image *a3079-_MG_71790000* from the *RAW-sRGB* dataset ($\times 4$).

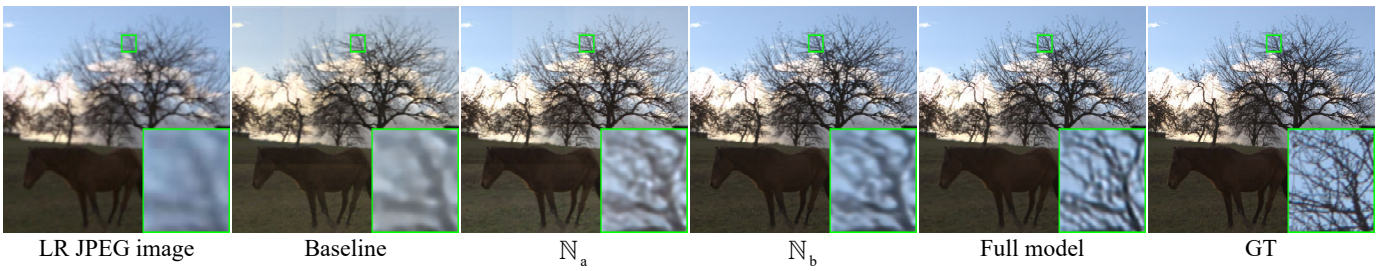


Fig. 9: Visual comparisons of the key modules in the PR sub-network ($\times 2$).

decrease in performance (though it is still better than the best sRGB image SR method shown in Table II, i.e., MIRNet-v2), which fully demonstrates the inferiority of sRGB image SR methods. On the other hand, although $\mathbb{N}_d$ outperforms $\mathbb{N}_c$, it is still inferior to the original PRNet, which indicates that it is inappropriate to extract guidance features from the RAW image. The qualitative results are given in Fig. 10. Though colors obtained by $\mathbb{N}_c$ and $\mathbb{N}_d$ are accurate, $\mathbb{N}_c$ produces blurred textures and $\mathbb{N}_d$ has obvious ringing artifacts.

Structure of CCB: To further verify the effectiveness of CCB, it is compared with other two feature modulation methods, including the normal spatial feature transform (SFT) block [23] and the cross-layer attention block (CAB) of the standard *Transformer* [58]. The quantitative comparison is provided in Table V, and visual results are provided in Fig. 11. The variant model that replaces CCB with SFT adopts normal 3×3 convolution layers to learn the scaling and shifting parameters, i.e., $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in Eq. (9). On the other hand, the variant model with CAB is obtained by substituting the normal multi-head attention (MHA) for the MAM in CCB. For a fair comparison, in MHA, the attention map is computed along the channel dimension rather than the spatial dimension. From

Fig. 10: Visual comparison among different processing structures (×2). The *ISP simulator* which cascades the IR sub-network with an LUT learning model [71] has blurred results and slight color shifts. Without the ACG generator, textures reconstructed by variant model $\mathbb{N}_c$ are blurred. Variant model $\mathbb{N}_d$ with exchanged inputs suffers from obvious ringing artifacts. One can zoom in for better viewing.

TABLE III: Ablation study for the modules in PR sub-network (×2). "LP" is short for the Laplacian pyramid decomposition.

Metric	Baseline	$\mathbb{N}_a$	$\mathbb{N}_b$	Full model
LP	×	✓	✓	✓
CLCM	×	×	✓	✓
CRM	×	×	×	✓
Parameters (M)	1.868	1.863	1.890	1.965
CPSNR (dB) ↑	22.09	22.48	31.11	**31.41**
SSIM ↑	0.7508	0.7817	0.8153	**0.8214**
LPIPS ↓	0.4389	0.3726	0.3352	**0.3194**

TABLE IV: Comparison of different processing structures (×2), where *ISP simulator* stands for cascading the IR sub-network with an LUT learning model [71]; $\mathbb{N}_c$ removes the ACG generator and accepts LR sRGB images as input; $\mathbb{N}_d$ is a variant model with the inputs of IR sub-network and ACG generator exchanged. All models have been adjusted to similar sizes.

Metric	ISP Simulator	$\mathbb{N}_c$	$\mathbb{N}_d$	PRNet
IR sub-network	✓	✓	✓	✓
PR sub-network	×	✓	✓	✓
ACG generator	×	×	✓	✓
Parameters (M)	1.938	1.948	1.928	1.965
CPSNR (dB) ↑	25.66	30.36	31.19	**31.41**
SSIM ↑	0.7703	0.7940	0.8182	**0.8214**
LPIPS ↓	0.3529	0.3718	0.3257	**0.3194**

Table V and Fig. 11, it can be observed that: 1) CAB achieves better performance than SFT, which verifies the effectiveness of the exploration of long-range dependency between the two inputs; 2) our MAM is superior to the other two methods, which suggests that modulating features with appropriate scaling and shifting parameters is more effective than simply conducting matrix multiplication between the learned attention map and the input features.

Number of Laplacian Pyramid Layers: The effects of different numbers of pyramid layers are compared in Table VI. Note that "$L = 1$" suggests that no decomposition is applied, and the features are directly refined by using CLCM and normal convolution layers. As can be seen, compared to "$L = 1$", the performance progressively improves as the number of decomposition layers increases. However, a slight performance drop can be observed in the case of "$L = 4$", which indicates that increasing the number of decomposition layers does not necessarily lead to better performance, as the extremely low resolution of the LF component may be harmful to the resolution-sensitive task. As a result, in the PR sub-network, "$L = 3$" is a suitable choice.

Fig. 11: Visual comparison among different modulation methods (×2). From left to right: SFT [23], CAB [58], CCB (ours) and GT.

TABLE V: Comparison of feature modulation methods (×2).

Metric	SFT [23]	CAB [58]	CCB
Parameters (M)	1.830	1.892	1.965
CPSNR (dB) ↑	31.17	31.28	**31.41**
SSIM ↑	0.8167	0.8193	**0.8214**
LPIPS ↓	0.3318	0.3247	**0.3194**

TABLE VI: Effects of the number of pyramid layers (×2).

Metric	L = 4	L = 3	L = 2	L = 1
Parameters (M)	2.047	1.965	1.934	1.963
CPSNR (dB) ↑	31.21	**31.41**	31.28	30.99
SSIM ↑	0.8188	**0.8214**	0.8199	0.8088
LPIPS ↓	0.3248	**0.3194**	0.3225	0.3350

Effectiveness of the Two-Stage Structure: Similar to [13], the proposed PRNet is also a two-stage structure, which first obtains a rough estimation by the IR sub-network and then restores details and corrects colors by the PR sub-network. To further evaluate the effectiveness of the two-stage framework, two variants are compared in Table VII, which are respectively named as $\mathbb{N}_e$ and $\mathbb{N}_f$ for simplicity. For the first variant $\mathbb{N}_e$, bicubic interpolation is used to build $\mathbf{X}_{\text{lin}}$ instead of the IR sub-network, while the structures of PR sub-network and the ACG generator are kept the same as before. For the second variant $\mathbb{N}_f$, the ACG generator is also kept the same as before, while the IR sub-network is replaced with multiple cascaded CLCMs, followed by a sub-pixel layer to directly reconstruct the HR sRGB image. Note that rearranged quarter-resolution image $\mathbf{X}_{\text{pack}}$ has a relatively small spatial resolution. Therefore, in variant $\mathbb{N}_f$, to maintain spatial information, the Laplacian pyramid is removed. For a fair comparison, we have adjusted the two variants, so that they have a model size similar to that of PRNet. By removing the IR sub-network, both variants become one-stage structures. As can be observed from Table VII, our PRNet significantly outperforms

TABLE VII: Comparison between the one-stage and two-stage structures ($\times 2$), where PRNet is a two-stage structure, while $\mathbb{N}_e$ and $\mathbb{N}_f$ are two variants with one-stage structures.

Metric	$\mathbb{N}_e$	$\mathbb{N}_f$	PRNet
Parameters (M)	1.968	1.916	1.965
CPSNR (dB) ↑	30.98	31.08	**31.41**
SSIM ↑	0.8112	0.8110	**0.8214**
LPIPS ↓	0.3372	0.3294	**0.3194**

TABLE VIII: Effects of the number of CCBs ($\times 2$), where m_{l} and m_{h} denote the numbers of CCBs for the CLCMs in the LF and the HF pyramid layers, respectively.

Metric	m_{l} = 3, m_{h} = 1	m_{l} = 4, m_{h} = 2	m_{l} = 5, m_{h} = 3
Parameters (M)	1.961	1.965	1.977
CPSNR (dB) ↑	31.29	31.41	**31.47**
SSIM ↑	0.8194	0.8214	**0.8217**
LPIPS ↓	0.3231	0.3194	**0.3167**
Running time (ms)	**14.74**	15.59	21.36

the two variants, which validates that the two-stage structure is more effective than directly learning sophisticated nonlinear mapping relations via the one-stage structure.

Effects of the Number of CCBs: Table VIII shows the effects of the number of CCBs in CLCM, where m_{l} and m_{h} denote the numbers of CCBs for the CLCMs in the LF and the HF pyramid layers, respectively. For a fair comparison, we have adjusted the channel dimensions of CCBs in different variants to achieve similar model sizes. It is clear that the more CCBs in CLCM, the better performance PRNet can achieve. However, from (m_{l} = 4, m_{h} = 2) to (m_{l} = 5, m_{h} = 3), only small performance increment can be obtained, while the running time of PRNet increases significantly. Therefore, we choose m_{l} = 4 and m_{h} = 2 in PRNet.

Structures of the IR Sub-Network and the ACG Generator: Four experiments are conducted to evaluate the structures of the IR sub-network and the ACG generator:

1) In the first experiment, the RCAGs in the IR sub-network and ACG generator are changed to the MHA of the standard transformer [58], and the results are shown in Table IX. Note that to maintain a similar model size with the original PRNet, the feature channels in the IR sub-network and ACG generator are reduced from 64 to 32. Similar to the proposed CCB, in this experiment, the attention map of MHA is computed across the channel dimension rather than the spatial dimension. As shown in Table IX, replacing RCAG with MHA results in a slight degradation in performance. Moreover, due to high computational complexity, additionally introducing MHA to the IR sub-network and ACG generator considerably increases the running time.
2) We evaluate the performance of a variant which replaces the RCABs in an RCAG with MV2 blocks [57], and the results are also given in Table IX. For a fair comparison, the model size of this variant has been adjusted to be similar to that of the original PRNet. As can be seen from Table IX, the variant using MV2 to extract and refine features suffers from performance drop, though a slightly faster speed can be achieved. Therefore, based on the first two experiments, compared with MHA and MV2, RCAG is more suitable for extracting features in the IR sub-network and ACG generator.
3) The effects of different numbers of RCAGs are evaluated in Table X, where s_1 and s_2 stand for the numbers of RCAGs in the IR sub-network and the ACG generator, respectively. For a fair comparison, the compared variants are adjusted to similar model sizes, and thus, increasing s_1 or s_2 requires decreasing the parameters in the PR sub-network, and vice verse. We can summarize from the shown results that: a) decreasing the number of RCAGs (s_1 = 1) in the IR sub-network results in worse performance, which suggests that a low-quality output of IR sub-network could mislead the following PR sub-network; b) increasing the number of RCAGs in the IR sub-network (s_1 = 3) also leads to a drop in performance, which is attributed to the fact that $\mathbf{X}_{\text{lin}}$ will be further enhanced by the PR sub-network, making a precise reconstruction of $\mathbf{X}_{\text{lin}}$ unnecessary; c) increasing the number of RCAGs (s_2 = 2) in the ACG generator reduces performance, owing to the ease of characterizing global color features, which do not require significant computational resources. Therefore, based on the above conclusions, we set $s_1 = 2$ and $s_2 = 1$ in our PRNet.
4) In the fourth experiment, we additionally evaluate the effects of applying dense connections in the ACG generator. From Table X, we are surprised to see that image quality slightly decreases when dense connections are further incorporated to the ACG generator. As have been corroborated by previous experiments, even with only one RCAG, the ACG generator is still able to well capture the ISP-pipeline-related color features. Therefore, it may not be essential to encourage feature reuse in this relatively shallow structure.

TABLE IX: Comparison among applying MV2, MHA, and RCAG in the IR sub-network and ACG generator ($\times 2$).

Method	MV2 [57]	MHA [58]	RCAG [32]
Parameters (M)	1.981	1.974	1.965
CPSNR (dB) ↑	31.33	31.37	**31.41**
SSIM ↑	0.8187	0.8198	**0.8214**
LPIPS ↓	0.3271	0.3249	**0.3194**
Running time (ms)	**13.96**	20.02	15.59

TABLE X: Effects of the number of RCAGs ($\times 2$), where s_1 and s_2 denote the numbers of RCAGs in the IR sub-network and the ACG generator, respectively; *w* and *w/o D* stand for with and without dense connections in the ACG generator, respectively.

Metric	$s_1 = 1$, $s_2 = 1$ (w/o D)	$s_1 = 2$, $s_2 = 1$ (w/o D)	$s_1 = 2$, $s_2 = 1$ (w D)	$s_1 = 3$, $s_2 = 1$ (w/o D)	$s_1 = 2$, $s_2 = 2$ (w D)
Parameters (M)	1.962	1.965	1.973	1.976	1.984
CPSNR (dB) ↑	31.30	**31.41**	31.37	31.24	31.28
SSIM ↑	0.8189	**0.8214**	0.8208	0.8195	0.8178
LPIPS ↓	0.3235	**0.3194**	0.3259	0.3227	0.3293

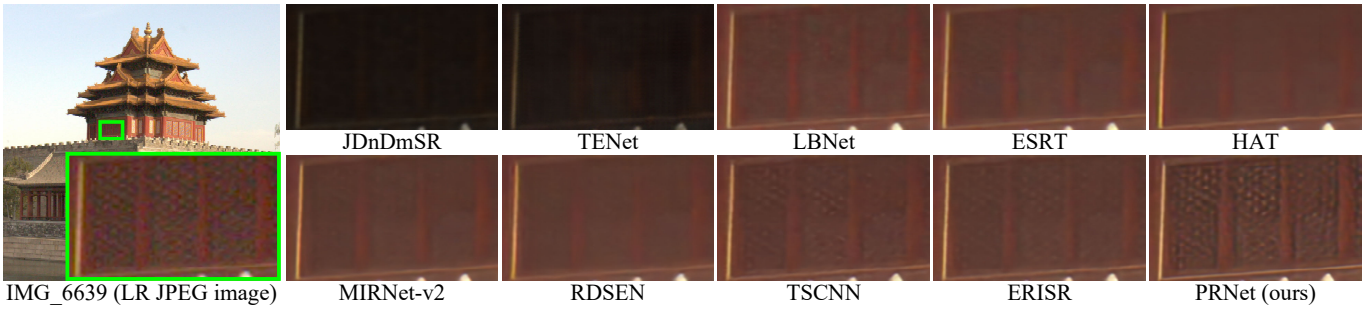


Fig. 12: Qualitative comparisons on image *6639* from the realistic dataset (×2).

Fig. 13: Visualization of the outputs of two sub-networks (×2). The GT HR linear/sRGB images are available in the *RAW-sRGB* dataset [21]. Images with different resolutions are resized to the same size for displaying. As GT HR linear images are not used in training, the intermediate results generated by IR sub-network have obvious color distortion. However, with the help of ACG generator, false colors can be well corrected by the following PR sub-network.

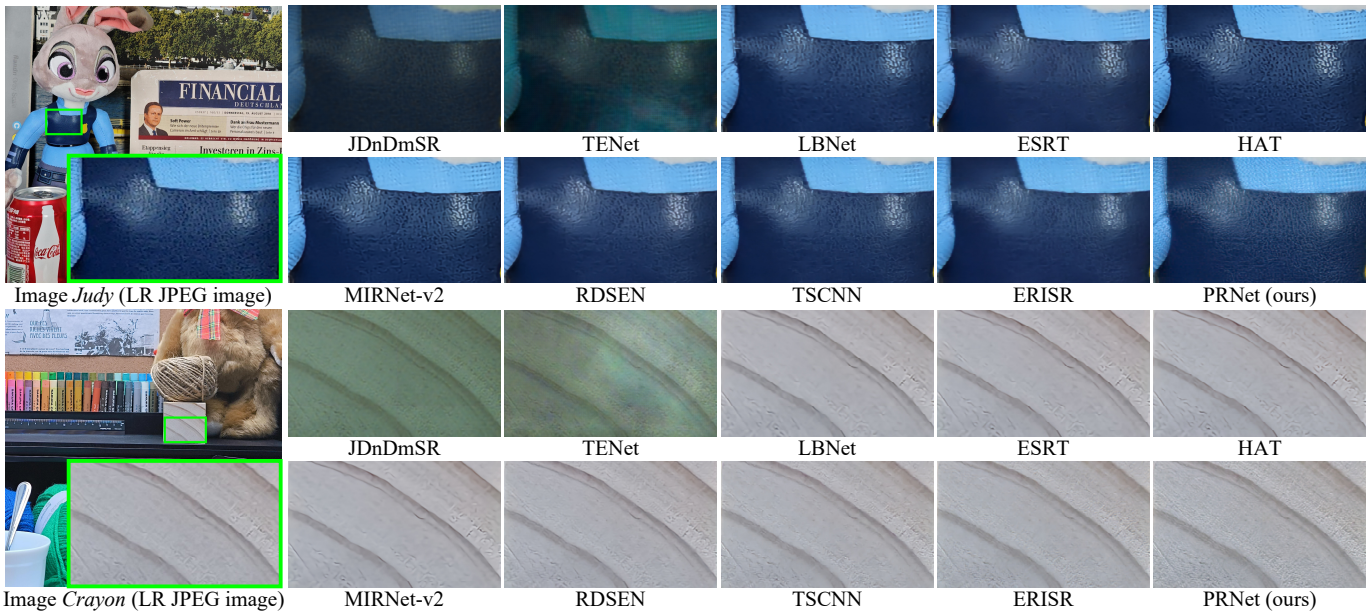


Fig. 14: Visual comparisons among different methods on images captured by unseen cameras (×2), where images *Judy* and *Crayon* are captured by Xiaomi 14 Pro and Xiaomi 13T Pro in a light box, respectively. One can zoom in for better viewing.

TABLE XI: Quantitative comparison among different methods on realistic dataset (×2). The best results are highlighted.

Method	NIQE ↓	PIQE ↓	NRQM ↑	PI ↓
LBNet [43]	7.46	84.72	3.13	7.17
ESRT [42]	7.18	84.68	3.12	7.03
HAT [47]	7.04	89.11	3.10	6.97
MIRNet-v2 [38]	7.11	91.08	3.10	7.01
RDSEN [18]	6.88	81.95	3.45	6.71
TSCNN [13]	6.78	80.41	3.55	6.62
ERISR [21]	**6.84**	74.79	3.60	6.62
PRNet (ours)	6.86	**72.45**	**3.88**	**6.49**

D. Experiments on the Realistic Dataset

The performance of different approaches is verified on the realistic dataset from [21] with a scale factor of ×2. An example of visual comparison is provided in Fig. 12. As can be seen, obvious color distortions occur in the results of JDnDmSR and TENet. Although colors in the results produced by the *sRGB-image-SR* methods (LBNet, ESRT, HAT, and MIRNet-v2) coincide with the LR sRGB images, small structures and fine details are lost. Owning to the utilization of RAW images, RDSEN, TSCNN, ERISR, and our PRNet produce much clearer results than the *sRGB-image-SR* methods. Particularly, our PRNet is able to preserve most details and the sharpest

TABLE XII: Comparisons of parameters, FLOPs and running time of different models ($\times 2$). The numbers in bold indicate the lightest/fastest methods. FLOPs and running time are measured on LR images with a resolution of 192×192.

Metric	JDnDmSR	TENet	LBNet	ESRT	HAT	MIRNet-v2	RDSEN	TSCNN	ERISR	PRNet (ours)
Parameters (M)	6.479	20.219	0.731	**0.678**	9.473	5.872	6.563	1.428	5.163	1.965
FLOPs (G)	260.00	466.62	213.21	123.87	475.63	317.43	237.93	85.45	93.83	**39.90**
Running Time (ms)	31.65	49.56	185.26	144.33	320.56	91.68	27.72	52.82	**9.66**	15.59

edges and also well suppress noises.

As GT HR images are not available in the realistic dataset, for quantitative experiment, we only evaluate four no-reference quality metrics, and the experimental results can be found in Table XI. Note that the results of JDnDmSR and TENet are excluded as they contain severe color distortion compared to the LR sRGB images. It can be found that our PRNet outperforms other competing methods in most cases, which is consistent with the visual results shown in Fig. 12.

E. Visualization of the Outputs of IR and PR Sub-networks

Fig. 13 visualizes the results of the IR and PR sub-networks. It can be observed that: 1) as the GT HR images in linear color space are not used in training, the rough estimation $\mathbf{X}_{\text{lin}}$ produced by the IR sub-network suffers from obvious color distortion; 2) with the guidance produced by the ACG generator, the following PR sub-network is still able to accurately restore vivid colors and fine details in the final results, which suggests that accurately reconstructing the intermediate results, i.e., $\mathbf{X}_{\text{lin}}$, is not necessary.

F. Generalization Capability of PRNet

In Section IV-D, we have demonstrated the effectiveness of PRNet on the realistic dataset [21]. However, as the crucial parameters of LUT and tone curves are withheld by camera manufacturers, the JPEG images produced by the Rawpy toolkit are not exactly in line with the JPEG images produced by the corresponding realistic camera ISP pipelines. Therefore, to further validate the generalization capability of our proposed method, we use the Xiaomi 13T Pro and Xiaomi 14 Pro smartphones to capture two light box scenes with different apertures, exposure times and ISO settings. The RAW and JPEG images produced by these two smartphones are fed to PRNet to produce the corresponding sRGB outputs with higher quality. Fig. 14 shows the visual comparisons among different methods. As can be seen, our PRNet not only preserves richer details than other methods, but also accurately renders vivid colors. These results well demonstrate the generalization ability of PRNet to other unseen camera ISP pipelines.

G. Model Size, Computational Burden, and Inference Time

To comprehensively compare the trade-off between image quality and model complexity, the average PSNR, required parameters, and running time of different methods are compared in Fig. 1. It is obvious that our PRNet achieves the best trade-off. Furthermore, Table XII lists the detailed model sizes, floating point operations per second (FLOPs)[2], and running time of different models. As can be seen, LBNet and ESRT have the fewest numbers of parameters. However, as MHA is applied and computed on spatial dimension, these two methods suffer from low running speeds. By ERISR, colors are corrected by simple operations, including matrix multiplications and additions, and thus a fast running speed can be guaranteed. Nevertheless, such simple operations are hard to imitate the complex ISP pipeline. Consequently, a larger model size is required by ERISR to obtain satisfactory results. On the contrary, due to the effective Laplacian pyramid decomposition, the successful incorporation of the guidance from the LR reference, and the proper exploration of inter-layer correlation among different pyramid layers, our PRNet is able to achieve superior performance with a relative low computational burden.

[2]FLOPs are measured by the Pytorch tool available at https://github.com/sovrasov/flops-counter.pytorch

V. Conclusion

In this paper, we proposed a new framework called PRNet to solve the RAW image SR problem. By this framework, the input RAW LR image is first fed to the IR sub-network to obtain an initial SR result in the linear color space. Simultaneously, external guidance is extracted from the LR sRGB image by the ACG generator. Afterwards, the PR sub-network decomposes the initial SR result, and then progressively refines the decomposed frequency components in a coarse-to-fine manner. To properly correct colors, CLCM is introduced to aggregate the external guidance and the input features in the LF layer of the PR sub-network. Furthermore, CRM in the HF layers is used to explore the inter-layer correlation, so as to obtain a faithful reconstruction. Through extensive experiments, we found that the Laplacian pyramid decomposition, the usage of the external guidance from the LR reference, and the exploration of inter-layer correlation among pyramid layers all contribute to the superior performance of PRNet. Compared with other SOTA methods, our PRNet is able to achieve a higher quality of reconstruction on both synthetic and realistic datasets at a low computational cost.

References

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[2] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 391–407.

[3] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Oct. 2021.

[4] S. W. Hasinoff *et al.*, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–12, Dec. 2016.

[5] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand, "Deep joint demosaicking and denoising," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–12, Dec. 2016.

[6] Y. Tan, K. Chang, H. Li, Z. Tang, and T. Qin, "Lightweight color image demosaicking with multi-core feature extraction," in *Proc. IEEE Conf. Vis. Commun. Image Process. (VCIP)*, Macau, China, Dec. 2020, pp. 136–139.

[7] L. Liu, X. Jia, J. Liu, and Q. Tian, "Joint demosaicing and denoising with self guidance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 2237–2246.

[8] T. Zhang, Y. Fu, and C. Li, "Deep spatial adaptive network for real image demosaicing," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Virtual, Feb. 2022, pp. 3326–3334.

[9] K. Feng *et al.*, "Mosaic convolution-attention network for demosaicing multispectral filter array images," *IEEE Trans. Comput. Imaging*, vol. 7, pp. 864–878, 2021.

[10] T. Brooks *et al.*, "Unprocessing images for learned raw denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 11 028–11 037.

[11] J. Liu *et al.*, "Learning raw image denoising with bayer pattern unification and bayer preserving augmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019, pp. 2070–2077.

[12] O. A. Elgendy, A. Gnanasambandam, S. H. Chan, and J. Ma, "Low-light demosaicking and denoising for small pixels using learned frequency selection," *IEEE Trans. Comput. Imaging*, vol. 7, pp. 137–150, 2021.

[13] K. Chang, H. Li, Y. Tan, P. L. K. Ding, and B. Li, "A two-stage convolutional neural network for joint demosaicking and super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4238–4254, Jul. 2022.

[14] H. C. Karaimer and M. S. Brown, "A software platform for manipulating the camera imaging pipeline," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 429–444.

[15] Y. Tang, K. Chang, M. Huang, and B. Li, "BMISP: Bidirectional mapping of image signal processing pipeline," *Signal Process.*, vol. 212, p. 109135, Nov. 2023.

[16] R. M. H. Nguyen and M. S. Brown, "RAW image reconstruction using a self-contained sRGB-JPEG image with only 64 KB overhead," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1655–1663.

[17] R. Zhou, R. Achanta, and S. Süsstrunk, "Deep residual network for joint demosaicing and super-resolution," in *Proc. Color Imaging Conf. (CIC)*, Vancouver, BC, Canada, Nov. 2018, pp. 75–80.

[18] X. Xu, Y. Ye, and X. Li, "Joint demosaicing and super-resolution (JDSR): Network design and perceptual optimization," *IEEE Trans. Comput. Imaging*, vol. 6, pp. 968–980, 2020.

[19] W. Xing and K. Egiazarian, "End-to-end learning for joint image demosaicing, denoising and super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Virtual, Jun. 2021, pp. 3506–3515.

[20] G. Qian *et al.*, "Rethinking learning-based demosaicing, denoising, and super-resolution pipeline," in *IEEE Int. Conf. Comput. Photography (ICCP)*, Pasadena, CA, USA, Aug. 2022, pp. 1–12.

[21] X. Xu, Y. Ma, W. Sun, and M.-H. Yang, "Exploiting raw images for real-scene super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1905–1921, Apr. 2022.

[22] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 1983.

[23] X. Wang, K. Yu, C. Dong, and C. Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 606–615.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[25] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1132–1140.

[26] X. Wang *et al.*, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, Munich, Germany, Sep. 2018, pp. 63–79.

[27] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 256–272.

[28] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4549–4557.

[29] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4809–4817.

[30] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 2472–2481.

[31] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 11 057–11 066.

[32] Y. Zhang *et al.*, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 294–310.

[33] J. Liu, J. Tang, and G. Wu, "Residual feature distillation network for lightweight image super-resolution," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, Virtual, Aug. 2020, pp. 41–55.

[34] Y. Mei *et al.*, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 5689–5698.

[35] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 527–542.

[36] K. Chang, M. Li, P. L. K. Ding, and B. Li, "Accurate single image super-resolution using multi-path wide-activated residual network," *Signal Process.*, vol. 172, p. 107567, 2020.

[37] M. Li *et al.*, "Multi-scale feature selection network for lightweight image super-resolution," *Neural Netw.*, vol. 169, pp. 352–364, 2024.

[38] S. W. Zamir *et al.*, "Learning enriched features for fast image restoration and enhancement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1934–1948, Feb. 2023.

[39] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Virtual, May 2021.

[40] J. Liang *et al.*, "SwinIR: Image restoration using swin transformer," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Montreal, BC, Canada, Oct. 2021, pp. 1833–1844.

[41] M. V. Conde, U. Choi, M. Burchi, and R. Timofte, "Swin2SR: SwinV2 transformer for compressed image super-resolution and restoration," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, Tel Aviv, Israel, Oct. 2022, pp. 669–687.

[42] Z. Lu *et al.*, "Transformer for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, New Orleans, LA, USA, Jun. 2022, pp. 456–465.

[43] G. Gao *et al.*, "Lightweight bimodal network for single-image super-resolution via symmetric CNN and recursive transformer," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Vienna, Austria, Jul. 2022, pp. 913–919.

[44] X. Zhang, H. Zeng, S. Guo, and L. Zhang, "Efficient long-range attention network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, Oct. 2022, pp. 649–667.

[45] S. W. Zamir *et al.*, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 5718–5729.

[46] F. Li *et al.*, "SRInpaintor: When super-resolution meets transformer for image inpainting," *IEEE Trans. Comput. Imaging*, vol. 8, pp. 743–758, 2022.

[47] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 22 367–22 377.

[48] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Colorado Springs, CO, USA, Jun. 2011, pp. 97–104.

[49] R. Tan, K. Zhang, W. Zuo, and L. Zhang, "Color image demosaicking via deep residual learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Hong Kong, China, Jul. 2017, pp. 793–798.

[50] K. Cui, Z. Jin, and E. Steinbach, "Color image demosaicking using a 3-stage convolutional neural network structure," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 2177–2181.

[51] T. Huang, F. F. Wu, W. Dong, G. Shi, and X. Li, "Lightweight deep residue learning for joint color image demosaicking and denoising," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Beijing, China, Aug. 2018, pp. 127–132.

[52] S. M. A. Sharif, R. Ali Naqvi, and M. Biswas, "Beyond joint demosaicking and denoising: An image processing pipeline for a pixel-bin image sensor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Virtual, Jun. 2021, pp. 233–242.
[53] J. Chen, S. Wen, and S.-H. G. Chan, "Joint demosaicking and denoising in the wild: The case of training under ground truth uncertainty," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Virtual, Feb. 2021, pp. 1018–1026.
[54] X. Zhang, Q. Chen, R. Ng, and V. Koltun, "Zoom to learn, learn to zoom," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3757–3765.
[55] X. Xu, Y. Ma, and W. Sun, "Towards real scene super-resolution with raw images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1723–1731.
[56] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1874–1883.
[57] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 4510–4520.
[58] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.
[59] M. Afifi, K. G. Derpanis, B. Ommer, and M. S. Brown, "Learning multi-scale photo exposure correction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Virtual, Jun. 2021, pp. 9157–9167.
[60] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1510–1519.
[61] Q. Wan, Z. Huang, J. Lu, G. Yu, and L. Zhang, "SeaFormer: Squeeze-enhanced axial transformer for mobile semantic segmentation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, May 2023.
[62] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Virtual, Apr. 2022.
[63] D. Menon and G. Calvagno, "Color image demosaicking: An overview," *Signal Process. Image Commun.*, vol. 26, no. 8, pp. 518–533, Oct. 2011.
[64] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[65] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 586–595.
[66] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.
[67] N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Proc. Nat. Conf. on Commun. (NCC)*, Mumbai, India, Feb. 2015, pp. 1–6.
[68] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Comput. Vis. Image Underst.*, vol. 158, pp. 1–16, May 2017.
[69] A. Ignatov *et al.*, "PIRM challenge on perceptual image enhancement on smartphones: Report," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, Munich, Germany, Sep. 2018, pp. 315–333.
[70] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, May 2019.
[71] H. Zeng, J. Cai, L. Li, Z. Cao, and L. Zhang, "Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2058–2073, Apr. 2022.

Mingyang Ling received the B.S. degree in communication engineering from Guangxi University, Nanning, China, in 2020. She is currently a Ph.D. student with School of Electrical Engineering, Guangxi University, Nanning, China. Her research interests include image super-resolution, demosaicking and denoising.

Kan Chang (Member, IEEE) received the B.S. degree in communication engineering, and the Ph.D. degree in communications and information systems from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2005 and 2010, respectively. From Feb. 2014 to Feb. 2015, he was a Visiting Scholar with the Department of Computer Science, Arizona State University (ASU), Tempe, AZ, USA. He is currently a Professor with the School of Computer and Electronic Information, and also with Guangxi Key Laboratory of Multimedia Communications and Network Technology, Guangxi University, Nanning, China. His research interests include image and video processing, etc. He has authored and co-authored over 70 scientific articles and has obtained 10 issued Chinese patents.

Mengyuan Huang received the B.S. degree in communication engineering from Southwest Jiaotong University, Chengdu, China, in 2020. She is currently pursuing her M.S. degree with School of Computer and Electronic Information, Guangxi University, Nanning, China. Her research interests include image enhancement and restoration.

Hengxin Li received the B.S. degree in communication engineering, and the M.S. degree in computer science from Guangxi University, Nanning, China, in 2018 and 2021, respectively. He joined Shenzhen Xiaomi Communications Co., Ltd in 2021, where he is currently a camera tuning engineer, responsible for noise reduction and image sharpening. His research interests include image super-resolution, demosaicking and denoising.

Shuping Dang (Senior Member, IEEE) received B.Eng (Hons) in Electrical and Electronic Engineering from the University of Manchester (with first class honors) and B.Eng in Electrical Engineering and Automation from Beijing Jiaotong University in 2014 via a joint '2+2' dual-degree program. He also received D.Phil in Engineering Science from University of Oxford in 2018. Dr. Dang joined in the R&D Center, Huanan Communication Co., Ltd. after graduating from University of Oxford and worked as a Postdoctoral Fellow with the Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST). He is currently a Lecturer with School of Electrical, Electronic and Mechanical Engineering, University of Bristol. The research interests of Dr. Dang include 6G communications and signal processing for communications.

Baoxin Li (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Maryland, College Park, in 2000. He joined Arizona State University (ASU) in 2004, where he is currently a Professor in Computer Science & Engineering and a Graduate Faculty Endorsed to Chair in the Computer Science, Electrical Engineering, and Computer Engineering programs. From 2000 to 2004, he was a senior researcher with SHARP Laboratories of America, Camas, WA, where he was the technical Lead in developing SHARP's HiIMPACT^{TM} Sports technologies. He was also an Adjunct Professor with the Portland State University from 2003 to 2004. His current research interests include computer vision and pattern recognition, image/video processing, multimedia, medical image processing, and statistical methods in visual computing. He won the SHARP Laboratories' President Award twice, in 2001 and 2004. He also received the SHARP Laboratories' Inventor of the Year Award in 2002. He received the National Science Foundation's CAREER Award from 2008 to 2009. He holds 16 issued US patents. Previously, he served as an Associate Editor for *IEEE Trans. on Image Processing* and *IEEE Trans. on Circuits and Systems for Video Technology*.