PROGRESSIVE PHYSICS-DRIVEN DEEP CONVERSION OF sRGB TO RAW IMAGES

Haiyan Shu, Zhengguo Li, Jinghong Zheng, and Zhuo Chen

Institute for Infocomm Research (I2R)
Agency for Science, Technology and Research (A*STAR), Singapore

ABSTRACT

Raw images are commonly processed by image signal processing (ISP) algorithms to standard RGB (sRGB) images in order to save the storage space and provide a suitable format for human visual system. Distortion and information loss are introduced in this procedure which degrades the performance of following tasks with these sRGB images as the input. To leverage the benefits from raw images, reversing sRGB images to raw images is expected. In this paper, a progressive physics-driven deep learning algorithm is proposed for the conversion of sRGB images to higher quality raw images by fusing physical-driven and data-driven deep learning approaches. Based on an observation that deep convolution neural networks (CNNs) are biased towards learning low-frequency functions, the proposed framework includes a high-frequency aware guidance branch to provide progressive guidance for the reconstruction branch. Experimental results indicate that the proposed algorithm improves reconstruction quality of existing data-driven approaches. It also reduces the sensitivity of existing data-driven approaches to test data in the sense that those images that are hard to be restored by the existing approaches are improved much more.

1. INTRODUCTION

With modern digital cameras, images are commonly captured by a monochromatic image sensor covered by a color filter array (usually a Bayer filter [1]), which arranges Red (R), Green (G) and Blue (B) color filters on a 2×2 matrix grid of image sensor. Therefore, output of the sensor is a mosaiced signal-channel raw image array with Bayer pattern. The raw image [2] has a high bit-depth (typically 12-14 bits) and can provide rich and high redundant information. In practical application, the captured unprocessed raw images can be converted to standard RGB (sRGB) images with three channels (Red, Green and Blue) by image signal processing (ISP) [3, 4, 5, 6, 7] algorithms in order to save the storage space and provide human visual system suitable images. Image signal processing (ISP) includes a sequence of operations such as demosaicing, noise reduction, white balance and image sharpening such that the generated image can be presented in sRGB format with high quality as possible.

The ISP process may introduce redundant data (by demosaicing), information loss (by color space conversion and gamma correction), problematic variations (by white balance), distortions (by noise reduction), and information loss to the image signal, which results in the degradation of task performance and redundant parameters for deep learning models with sRGB images as input. Due to the well preserved linear relationship with incident scene irradiance, an unprocessed raw image is more suitable than an sRGB image for a wide range computer vision tasks. How to reconstruct the near raw image without explicitly storing the original unprocessed raw image attracts the interest of research for object detection, classification, and recognition purpose.

Since ISP process is non-reversible, converting an sRGB image back to a raw image is an ill-posed inverse problem. In [8], a generic camera ISP process with invertible steps was proposed and sequentially inverting image processing transformations are applied to reconstruct the synthetic raw image. In these processing steps, specific camera parameters are required which are generally not available for sRGB image dataset. Raw image reconstruction based on metadata can restore the sRGB image back to its corresponding raw value [9, 10]. This approach utilized the metadata to parameterize the mapping function from an sRGB image to an unprocessed raw image. Additional overhead to store specific metadata is needed. Recently, data-driven method, which is based on deep learning framework, was well studied [11, 12, 13, 14] for an sRGB image to a raw image reconstruction. Without analytical solution based on physical models, deep learning based approach relies on big data to provide an end-to-end solution which is an effective way to solve the problem. In [11], learnable histogram layer was introduced to extract both the global and the local color histogram as the feature. The entire pipeline is to incorporate the learnable histogram layer into conventional Convolutional neural network (CNN). CycleISP [12] introduced multiple dual attention blocks (DABs) to identify the more useful features by channel attention and spatial attention mechanisms. With the DAB, less useful features are suppressed to address the attention awareness. Recursive residual groups (RRGs) are then to group DABs to extract deep features to form the deep CNN system for sRGB images to raw image conversion. Converting sRGB image to another device-independent color space, CIE XYZ [15] which is also linear with respect to scene irradiance, was studied in [13]. In this framework, global processed patch was derived by the predicted local residue image from a fully CNN network. The global processed patches were then fed into another sub-network to generate a linear image state (i.e. CIE XYZ format). In [14], an invertible ISP (InvISP) framework was presented. Leverage the affine coupling layers, InvISP framework has the inherent reversibility where visually appealing sRGB images were rendered in the forward path and the near raw sensor data can be recovered through the inverse process.

Though data-driven deep learning approaches provide promising sRGB to raw solution, they have the limitation that is biased towards learning low-frequency information. To circumvent such shortcoming, we propose to use a guided filter [16, 17] to extract the high-frequency information from the sRGB image, and utilize this extracted information to guide the inference procedure for raw reconstruction which is a deep CNN based system. Since the nature of the convolution operator makes it hard to model long range interactions, the physics-driven guidance is incorporated into the raw reconstruction path progressively. In addition, the DABs [12]

are adopted to leverage their capability to suppress the less useful features and enhance the more useful features in the reconstruction process. In the proposed high frequency guided network (HFG-Net), the high-frequency information extracted from sRGB images by the physics-driven method is integrated with the deep learning based approach to recover the unknown high-frequency information for the raw image reconstruction. Clearly, the HFG-Net includes a promising idea on restoring the high-frequency information for deep-learning based methods, and reducing the sensitivity of these methods to the test data.

The rest of this paper is organized as follows. The guided filter used for the high-frequency information extraction is presented in Sec. 2. Details on the proposed HFG-Net are presented in Sec. 3. Experimental results are provided in Sec. 4 and conclusion remarks are listed in Sec. 5.

2. PRELIMINARY ON GUIDED FILTER

Guided image filters such as [16, 17] are edge-preserving filters to retain the sharp edges and detailed information in the image. The noise or texture information is expected to be removed in this filtering process. This type of filters can be used to guide the learning procedure for different image reconstruction applications[18, 19, 20]. The function of these filters is to decompose an image S into two parts as follows:

$$S(p) = S_b(p) + S_e(p), \tag{1}$$

where S_b is a reconstructed image formed by homogeneous regions with sharp edges, and S_e is high-frequency information such as noise or texture, and $p(=(x,y))$ is a pixel position.

Let $\Omega_\zeta(p')$ be a square window centered at the pixel p' of a radius ζ. The homogeneous image with sharp edges S_b can be extracted with a linear transform from the original image S and the guidance image $G(p)$ in the window $\Omega_\zeta(p')$ [16]:

$$S_b(p) = a_{p'}G(p) + b_{p'}, \forall p \in \Omega_\zeta(p'), \tag{2}$$

where $a_{p'}$ and $b_{p'}$ are two constants in the window $\Omega_\zeta(p')$.

To obtain $a_{p'}$ and $b_{p'}$ in (2), the following quadratic optimization problem is to be solved [17]:

$$\arg\min_{a_{p'},b_{p'}} \{\sum_{p\in\Omega_\zeta(p')} [\Gamma^G_{p'}(S_b(p) - S(p))^2 + \lambda a^2_{p'}]\}, \tag{3}$$

where λ is a regularization parameter and $\Gamma^G_{p'}$ is the edge-aware weighting which is defined as [17]

$$\Gamma^G_{p'} = \frac{1}{N}\sum_{p=1}^{N} \frac{\sigma^2_{G,1}(p') + \varepsilon}{\sigma^2_{G,1}(p) + \varepsilon}. \tag{4}$$

Here, N is the total pixel number in the guidance image G, $\sigma^2_{G,1}(p')$ is the variance of guidance image G in the window $\Omega_1(p')$, and ε is a small constant. For an 8-bit image in the range [0, 255], ε takes value of 1.

From the above, the optimal values of $a_{p'}$ and $b_{p'}$ are computed as

$$\begin{cases} a^*_{p'} = \frac{\Gamma^G_{p'}\mathrm{cov}_{S,G,\zeta}(p')}{\Gamma^G_{p'}\sigma^2_{G,\zeta}(p')+\lambda} \\ b^*_{p'} = \mu_{S,\zeta}(p') - a^*_{p'}\mu_{G,\zeta}(p') \end{cases}, \tag{5}$$

where $\mathrm{cov}_{S,G,\zeta}(p')$ is $(\mu_{G\odot S,\zeta}(p') - \mu_{G,\zeta}(p')\mu_{S,\zeta}(p'))$, $\odot$ is the Hadamard product. $\mu_{G,\zeta}(p')$, $\mu_{S,\zeta}(p')$, and $\mu_{G\odot S,\zeta}(p')$ are the mean values of G, S, and $G \odot S$ in the window $\Omega_\zeta(p')$, respectively.

When the images S and G are the same, the optimal values of $a_{p'}$ and $b_{p'}$ are computed as

$$\begin{cases} a^*_{p'} = \frac{\Gamma^S_{p'}\sigma^2_{S,\zeta}(p')}{\Gamma^S_{p'}\sigma^2_{S,\zeta}(p')+\lambda} \\ b^*_{p'} = \mu_{S,\zeta}(p') - a^*_{p'}\mu_{S,\zeta}(p') \end{cases}. \tag{6}$$

The reconstructed image formed by homogeneous regions with sharp edges is finally produced as

$$S_b(p) = \bar{a}_p G(p) + \bar{b}_p, \tag{7}$$

where $\bar{a}_p$ and $\bar{b}_p$ are computed as [19]

$$\{\bar{a}_p, \bar{b}_p\} = \frac{1}{|\Omega_\zeta(p')|}\sum_{p'\in\Omega_\zeta(p)} \{a^*_{p'}, b^*_{p'}\}, \tag{8}$$

and $|\Omega_\zeta(p')|$ is the cardinality of the set $\Omega_\zeta(p')$.

3. THE PROPOSED sRGB TO RAW CONVERSION

Deep learning approach is commonly used to provide an end-to-end solution to reconstruct the raw images from the sRGB images. The proposed high frequency guided network (HFG-Net) is proposed on top of the conventional convolutional neural network by fusing physical features which are not sensitive to data.

In the CycleISP [12], the DAB was introduced to distinguish the features with more useful information. The less useful information is suppressed by channel attention and spatial attention mechanisms. The DAB treats different features and pixels unequally, which can provide additional flexibility in dealing with different types of information. Each DAB combines a channel attention block and a spatial attention block in channel-wise and pixel-wise features where the architecture is shown in Fig. 1. The proposed HFG-Net adopts this approach to build the reconstruction branch as shown in Fig. 2. In this framework, reconstruction branch is composed of several recursive residual groups (RRGs) as shown in Fig. 3. Each RRG contains multiple DABs. The performance of reconstruction branch is enhanced with more useful information distinguished by the DABs.

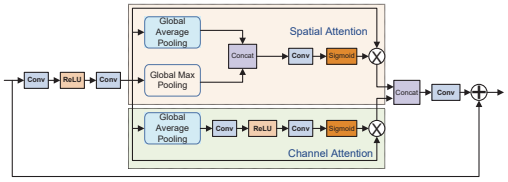


Fig. 1: *Dual Attention Block.*

It was pointed out in [21] that the deep CNNs are biased towards learning low-frequency functions. Inspired by an exposure aware guidance branch (EGB) [22], we explore the utilization of the guided filter [16] to extract the high frequency information from the sRGB image S_{rgb} and utilize the high frequency information to guide the reconstruction branch for the raw image reconstruction. With this information, a high-frequency aware guidance

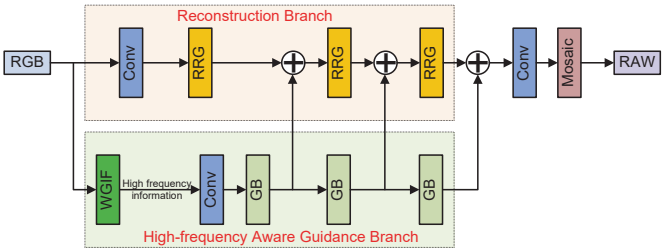


Fig. 2: *The proposed HFG-Net for converting an sRGB image into a near raw image is composed of a reconstruction branch (RB) and a high-frequency aware guidance branch (HFGB). The HFGB helps the GB to preserve detailed information.*

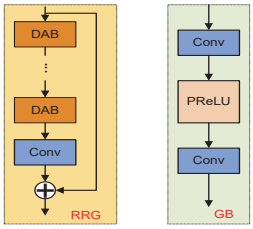
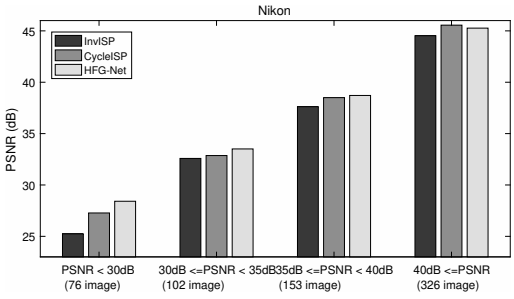
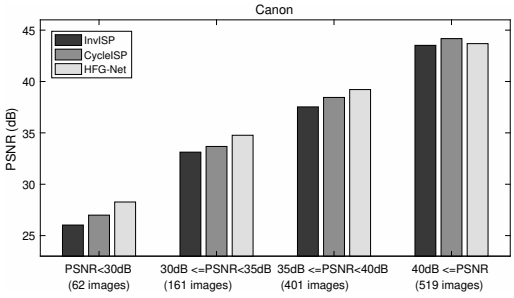


Fig. 3: *Recursive residual group (RRG) and Guide Block (GB).*



(a) Nikon



(b) Canon

Fig. 4: *Comparison of image recovering with different quality.*

branch is formed in the proposed HFG-Net as in Fig. 2. The high-frequency information is used to remind the reconstruction branch to pay more attention to its weakness and help the restoration of the unknown high-frequency information in the reconstructed raw images. This is somewhat similar to the idea in [23].

In the high-frequency aware guidance branch, each guide block (GB) is formed as in Fig. 3. Since the nature of the convolution operator makes it hard to model long range interactions, the physics-driven guidance is integrated into the reconstruction branch progressively. Features of the guidance branch are combined with those of the reconstruction branch in a multi-stage manner such that the high-frequency information is compensated in each step. Subsequently, the quality of the reconstructed raw image will be improved, and the sensitivity of the reconstruction branch with respect to data is also reduced.

Besides the structure, loss functions are also important for the HFG-Net. Let the ground-truth image of S_{raw} is denoted as $\hat{S}_{raw}$. Two loss functions are defined as

$$\begin{cases} L_r = \|S_{raw} - \hat{S}_{raw}\|_1 \\ L_h = \|\psi_h(S_{raw}) - \psi_h(\hat{S}_{raw})\|_1 \end{cases}, \tag{9}$$

where the function $\psi_h(S)$ is defined as $(S - \mu_{S,1})$, and $\mu_{S,1}(p)$ is the mean value of S in a 3×3 window with the pixel p as the center. The total loss is computed by

$$L_T = L_r + L_h. \tag{10}$$

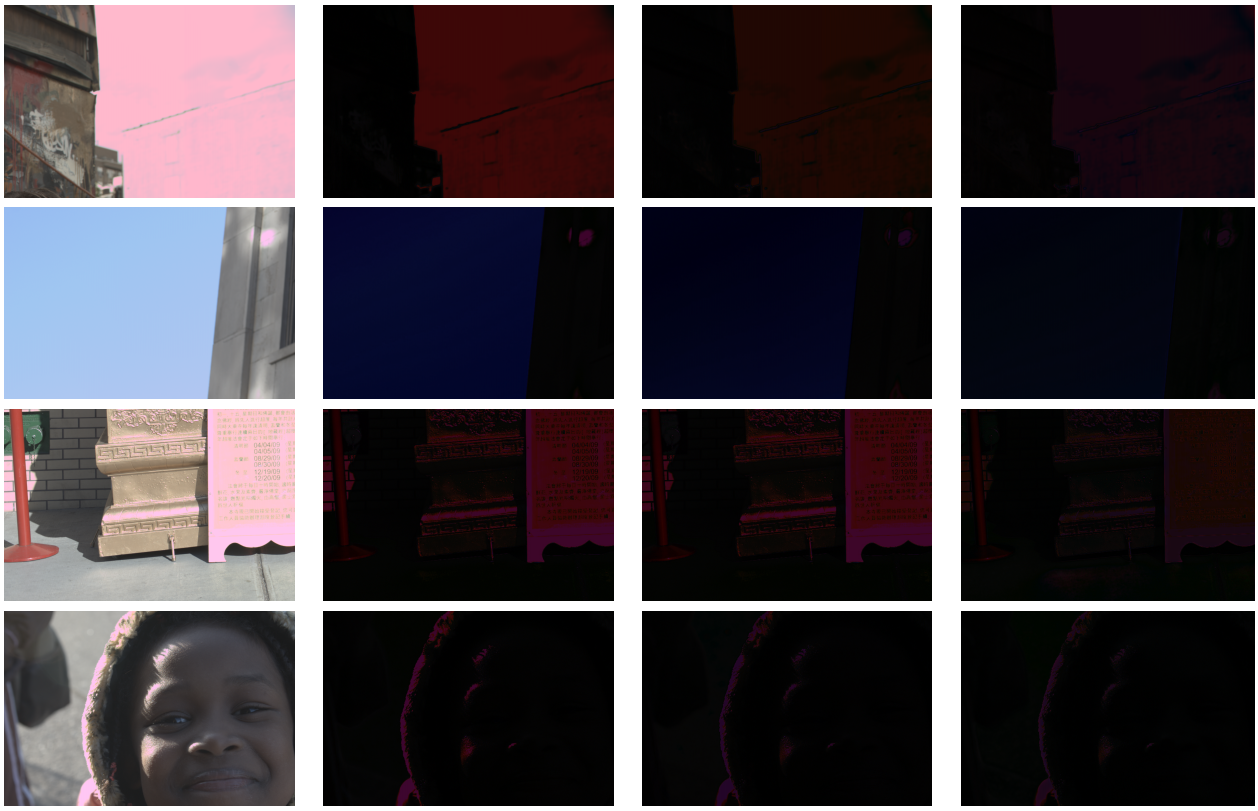
4. EXPERIMENTAL RESULTS

The dataset in [14] is adopted in our experiments. The dataset comes from MIT-Adobe FiveK dataset [24]. There are two sets which are captured by using the Canon EOS 5D and the Nikon D700 respectively. Training and testing data are randomly partitioned with 650/127 images for the Canon set and 414/72 images for the Nikon set.

Raw images are preprocessed by using the white balance parameters provided by camera metadata which follows the same steps in [14]. The evaluation is based on the preprocessed raw images to eliminate the camera dependent parameters. The ground-truth sRGB images are generated from the raw images using the Libraw library [25] to simulate the ISP procedure which is the most commonly used approach in a wide range digital cameras. The sRGB images are saved in the JPEG format which is the most popular image format in computer vision dataset. In the testing, each image is split into 9 sub-images. So the total number of testing images is 657 for the Nikon data set and 1143 for the Canon data set.

Our proposed approach is compared with the invertible ISP (InvISP) [14] and cycleISP [12] which are two well known reverse ISP approaches. The InvISP adopts an affine coupling structure where the ISP and reverse ISP share the same framework and parameters. The CycleISP includes the DABs on top of conventional CNN which presents promising results. Quantitative comparison results are presented in Table. 1. Obviously, the CycleISP outperforms the InvISP in terms of both PSNR and SSIM. Furthermore, the proposed HFG-Net presents a better output over the CycleISP in terms of PSNR. Average improvements of 0.16dB and 0.33dB are achieved for Nikon and Canon subsets, respectively. The CycleISP and the proposed HFG-Net are comparable from the SSIM point of view. This result shows that dual attention blocks can well extract the useful information for the following reconstruction. And the proposed progressive high-frequency information enhancement well improves the reconstruction with the edge and high frequency information.

Existing data-driven approaches are usually sensitive to the test data. To compare the sensitivity of different approaches, each data set is partitioned into four subsets based on the reconstructed



(A) (B) (C) (D)

Fig. 5: *The qualitative comparison of different reverse ISP approaches. (A) Ground-truth raw images and error map of different approaches (B) InvISP (C) CycleISP, and (D) Proposed HFG-Net.*

Table 1: *Quantitative evaluation among our model and other approaches.*

Methods	Nikon		Canon	
	PSNR (dB)	SSIM	PSNR (dB)	SSIM
InvISP	38.625	0.9602	39.020	0.9495
CycleISP	39.793	0.9681	39.683	0.9709
HFG-Net	39.954	0.9687	40.015	0.9704

PSNR by the InvISP: less than 30dB (S1), between 30dB to 35dB (S2), between 35db to 40dB (S3), and higher than 40dB (S4). The comparison results based on these subsets are shown in Fig. 4. The HFG-Net outperforms the InvISP in all the subsets. A significant improvement can be observed in the subset S1 where 3.2dB and 2.3dB are achieved for the Nikon and Canon sets, respectively. The HFG-Net also outperforms the CycleISP for the subsets of S1, S2, and S3, and is comparable to the CycleISP for the subset S4. In the subset S1, the HFG-Net achieves 1.14dB and 1.28dB improvement over the CycleISP for Nikon and Canon sets, respectively. This confirms the capability of the proposed HFG-Net in recovering unknown high frequency information that is hard to be restored by other approaches and reducing the sensitivity to the test data.

Besides the objective comparison, the three different methods are also compared subjectively by presenting the error map over the unprocessed raw in Fig. 5. Fig. 5 (B) to (D) are the results for the InvISP, CycleISP, and the proposed HFG-Net. Clearly, the reconstruction error for the InvISP is very clear. At the same time, the superiority of the HFG-Net over the CycleISP is also visible.

5. CONCLUSION

A novel high frequency information guided network is proposed for the conversion of sRGB images to raw images. The framework is built on top of convolution neural network (CNN) with deep features extracted by grouped dual attention blocks (DABs) as input. To overcome the limitation of CNN that is biased towards learning low-frequency functions, a high-frequency aware guidance branch is introduced for the reconstruction. This physics-driven guidance branch is incorporated into the raw reconstruction progressively to model long range interactions for the raw images recovering. Experimental results shows that the proposed HFG-Net significantly improves the reconstruction quality in terms of PSNR, and reduces the sensitivity of the data-driven approach to the test data. Clearly, the proposed progressive physics-driven guidance approach can also be extended to study other CNN-based low level image processing problems.

Acknowledgement

This research is supported in part by A*STAR under AHSF project No. C211118005.

References

[1] B. E. Bayer, "Color imaging array," United States Patent 3,971,065, 1976.

[2] B. Fraser, "Understanding digital raw capture," Digital camera raw file support, Adobe Systems Incorporated, vol. 48, 2004.

[3] R. Ramanath, W. E. Snyder, Y. Yoo, and M. S. Drew, "Color image processing pipeline," IEEE Signal Processing Magazine, vol. 22, no. 1, pp. 34–43, 2005.

[4] E. Schwartz, R. Giryes, and A. M. Bronstein, "Deepisp: Toward learning an end-to-end image processing pipeline," IEEE Transactions on Image Processing, vol. 28, no. 2, pp. 912–923, 2018.

[5] P. Hansen, A. Vilkin, Y. Krustalev, J. Imber, D. Talagala, D. Hanwell, M. Mattina, and P. N. Whatmough, "Isp4ml: The role of image signal processing in efficient deep learning vision systems," in 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 2438–2445.

[6] S. A Sharif, R. A. Naqvi, and M. Biswas, "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, 2021, pp. 233–242.

[7] M. V. Conde, S. McDonagh, M. Maggioni, A. Leonardis, and E. Pérez-Pellitero, "Model-based image signal processors via learnable dictionaries," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 1, 2022, pp. 481–489.

[8] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, and J. T. Barron, "Unprocessing images for learned raw denoising," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11 036–11 045.

[9] R. M. Nguyen and M. S. Brown, "Raw image reconstruction using a self-contained srgb-jpeg image with only 64 kb overhead," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1655–1663.

[10] S. Nam, A. Punnappurath, M. A. Brubaker, and M. S. Brown, "Learning srgb-to-raw-rgb de-rendering with content-aware metadata," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17 704–17 713.

[11] S. Nam and S. Joo Kim, "Modelling the scene dependent imaging in cameras with a deep neural network," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1717–1725.

[12] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Cycleisp: Real image restoration via improved data synthesis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2696–2705.

[13] M. Afifi, A. Abdelhamed, A. Abuolaim, A. Punnappurath, and M. S. Brown, "Cie xyz net: Unprocessing images for low-level computer vision tasks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 9, pp. 4688–4700, 2021.

[14] Y. Xing, Z. Qian, and Q. Chen, "Invertible image signal processing," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6287–6296.

[15] T. Smith and J. Guild, "The cie colorimetric standards and their use," Transactions of the optical society, vol. 33, no. 3, p. 73, 1931.

[16] K. He, J. Sun, and X. Tang, "Guided image filtering," in European conference on computer vision. Springer, 2010, pp. 1–14.

[17] Z. Li, J. Zheng, Z. Zhu, W. Yao, and S. Wu, "Weighted guided image filtering," IEEE Transactions on Image processing, vol. 24, no. 1, pp. 120–129, 2014.

[18] Z. Li and J. Zheng, "Single image de-hazing using globally guided image filtering," IEEE Transactions on Image Processing, vol. 27, no. 1, pp. 442–450, 2017.

[19] Z. Li, J. Zheng, and J. Senthilnath, "Simultaneous smoothing and sharpening using iwgif," in 2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022, pp. 861–865.

[20] C. Zheng, Y. Li, and S. Wu, "Single image deraining via rain-steaks aware deep convolutional neural network," arXiv preprint arXiv:2209.07808, 2022.

[21] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in International Conference on Machine Learning. PMLR, 2019, pp. 5301–5310.

[22] Y. Xu, Z. Liu, X. Wu, W. Chen, C. Wen, and Z. Li, "Deep joint demosaicing and high dynamic range imaging within a single shot," IEEE Transactions on Circuits and Systems for Video Technology, 2021.

[23] Z. Li, W. Gao, A. H. Adiwahono, and W. Y. Yau, "Hierarchical random exploring with multiple linking modes," in TENCON 2017-2017 IEEE Region 10 Conference. IEEE, 2017, pp. 2104–2109.

[24] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in CVPR 2011. IEEE, 2011, pp. 97–104.

[25] "Libraw: raw images decoder library." [Online]. Available: www.libraw.org