2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10223129

FLASH COMPENSATED LOW-LIGHT ENHANCEMENT VIA HIERARCHICAL NETWORK PREDICTION

Haowei Kuang[1], *Haofeng Huang*[1], *Wenhan Yang*[2], *Jiaying Liu**[1]

[1]Wangxuan Institute of Computer Technology, Peking University, Beijing, China
[2]Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

Photography in low-light conditions suffers from dense noise and insufficient light. Flash photography, introducing extra light sources, performs better at suppressing noise and revealing details, while being interrupted by unnatural ambient illumination. This paper offers an analysis of the pros and cons to utilize low-light and flash images for enhancement, which inspires us to design a unified sample-adaptive CNN to capture diverse focuses from different inputs in a complementary way. Specifically, a Flash Compensated Dynamic Filtering Network is proposed to utilize the revealed details of flash images to compensate for fine structure reconstruction in low-light enhancement. To adaptively fuse information from misaligned low-light and flash image pairs, our network is designed with three distinctive features. Firstly, we adopt a layer-wise regression strategy, where results are predicted from the single input first and then fused to sufficiently leverage complementary information. Secondly, we employ a sample-adaptive mechanism, where each pixel is estimated with its distinctive parameters augmented by weighted residual connections. Finally, we utilize a coarse-to-fine architecture, where features are extracted by diversified receptive fields to utilize hierarchical contextual information. Experimental results demonstrate that the three design principles lead to the significant superiority of the proposed method over state-of-the-art methods.

Index Terms— Low-Light Enhancement, Flash Compensation, Dynamic Filtering Network, Kernel Prediction.

1. INTRODUCTION

Photography in low-light environments causes inevitable image degradation, mainly including illumination distortion and intensive noise. In recent decades, researchers have been continuously investigating the task of low-light image enhancement. Classical methods for low-light image enhancement develop like adjusting global illumination level via Gamma transform or histogram equalization [1]. Later, some handcrafted design methods based on retinex theory are proposed. More recently, the presence of deep learning theory promotes rapid development of low-light enhancement and significantly improves the performance [2, 3]. Although many methods can reconstruct satisfactory global illumination, it is still challenging to restore the fine structure of low-light images due to inevitable information loss caused by insufficient light. Flash photography mitigates the information loss by adding artificial light to illuminate the foreground and improve the signal-to-noise ratio of photos. However, most flashes are point lights, and therefore seriously damage the natural ambient illumination.

A series of methods are developed to introduce flash images to compensate for the fine structure of the enhanced low-light images with the revealed details in flash images. In [4], Deng *et al.* built a Common and Unique Information Splitting Network (CU-Net) to split the common information shared among different modalities for image restoration. Liu *et al.* [5] proposed Frequency-relevant Residual Learning (FRL) to regress the denoising result by integrating the frequency-domain information from different modalities. Li *et al.* [6] proposed Deep convolutional Joint image Filtering (DJF) to selectively transfer salient structures. Meanwhile, other methods based on location-specific kernels prediction such as Deformable Kernel Networks (DKN) [7] and deep denoising method with Flash/No-Flash pairs (deepFnF) [8] also achieve ideal image reconstruction. However, two aspects are not adequately considered. Firstly, the information of low light and flash images are simply mixed, without taking full advantage of their complementary characteristics. Secondly, low-light distortion is sample-dependent while most methods adopt a network with fixed weights, or just predict one-layer adaptive kernels.

To address the aforementioned issues, we propose a Flash Compensated Dynamic Filtering Network to restore ideal environmental illumination and fine structures of low-light images by using flash image compensation. The results are predicted from the single input firstly and then fused to sufficiently leverage complementary information. Our network consists of a couple of Sample-Adaptive Element-wise Con-

* Corresponding Author. This work was supported by the National Natural Science Foundation of China under Contract No.62172020, and a research achievement of Key Laboratory of Science, Techonology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). This research work was also partially supported by the Basic and Frontier Research Project of PCL and the Major Key Project of PCL.

volutional Neural Network (SAECNN) for low-light and flash image restoration. It generates sample-adaptive parameters for each pixel, benefiting the modeling of diverse illumination and noise. Unlike traditional dynamic filtering networks [9] which only employ a simple one-layer filtering structure, our network predicts a filtering network with multi-layer kernels and weighted residual connections, providing diversified receptive fields and the ability to extract hierarchical contextual information. Experiments demonstrate the effectiveness of flash compensation and the superiority of our network design.

The rest of the paper is organized as follows. Sec. 2 introduces the proposed method, where we first describe our motivation then the network architecture and loss function are presented. Experimental results are shown in Sec. 3 and concluding remarks are given in Sec. 4.

2. FLASH COMPENSATED DYNAMIC FILTERING NETWORK

2.1. Degradation Analysis and Motivation

We start with the retinex theory, which provides a clear view to review degradation in flash photography. The images taken under normal light I_{GT} can be represented as:

$$I_{GT} = R \cdot L_{GT}, \tag{1}$$

where R is the reflectance of the scene and L_{GT} is the normal illumination layer. In the low-light condition, we can model the low-light image I and the flash image I_f as:

$$\begin{aligned} I &= R \cdot L + N, \\ I_f &= R \cdot (L + L_o) + N_f, \end{aligned} \tag{2}$$

where L is the illumination layer in low-light and N, N_f are the noise of low-light and flash images. Here, we separate the illumination layer under flash into ambient light L and flash L_o.

It is noted that, for low-light illumination distortion, the transformation from normal-light illumination L_{GT} to low-light illumination L can be approximately regarded as proportionally dimming after linearization [8]. Therefore, we can connect the normal-light and low-light scenarios by:

$$\begin{aligned} I &= R \cdot L_{GT}^{\gamma} + N, \\ I_f &= R \cdot (L_{GT}^{\gamma} + L_o) + N_f, \end{aligned} \tag{3}$$

where γ defines the intensity of the Gamma transformation. Therefore, we obtain the observation of I and I_f, and aim to restore $I_{GT} = R \cdot L_{GT}$.

By comparing the two formulas in Eqn. (3), we can clearly see their advantages and disadvantages for restoration:

- I_f benefits noise removal and detail revealing. Brighter images are less disturbed by noise [10], N_f is much less intensive than N. Therefore, flash images I_f can better reveal details.

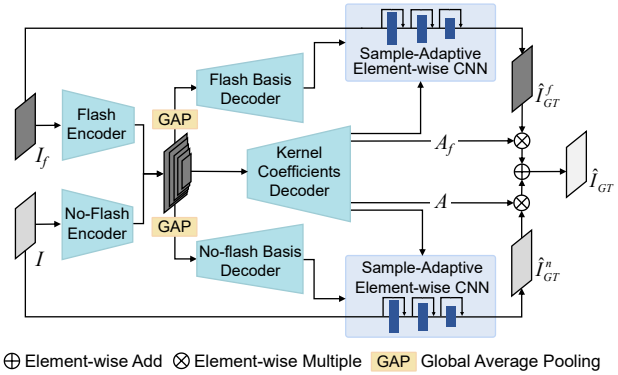

Fig. 1. Model architecture of our proposed network. The network predicts two Sample-Adaptive Element-wise CNNs for image reconstruction and fuses results to output the final enhanced results.

- I benefits illumination restoration. I_f introduces additional degradation in illumination due to the presence of L_o. L_o is entirely generated by the unnatural flash light, irrelevant with expected normal light illumination.

In our work, we aim to take a unified framework to make full use of I and I_f complementarily. More in detail, sample-adaptive CNN is designed to operate on both I and I_f. The module is flexible to deal with degradation dynamically. It estimates $\hat{I}_{GT}^{n}$ from I to restore better illumination and $\hat{I}_{GT}^{f}$ to obtain more details from I_f. $\hat{I}_{GT}^{n}$ and $\hat{I}_{GT}^{f}$ are finally fused to estimate I_{GT}, getting their benefit and keeping away from the influence of L_o and N.

2.2. Network Architecture

The network follows an encoder-decoder architecture with skip connections [11] , which includes two encoders and three decoders. The outputs of decoders predict the weights of a couple of Sample-Adaptive Element-wise Convolutional Neural Networks and importance maps, as shown in Fig. 1. Flash and low-light images are filtered by the predicted networks separately and fused guided by the importance maps to reconstruct the final output.

Predicting CNN kernels. To reduce memory burden and computational costs for Sample-Adaptive Element-wise CNN prediction, we inherit [12] which estimates kernels by predicting a set of low-rank kernel basis and related per-pixel coefficient vectors.

To sufficiently leverage complementary information from the flash and low-light images, we predict two distinct sets of per-pixel coefficient elements, kernel biases, residual connection weights and importance maps by one decoder, and two low-dimensional basis by the other two decoders identical in structures without weight sharing. Skip connections to basis decoders are global average pooled. The k_{th} kernel of Sample-Adaptive Element-wise CNN at pixel r is computed

by:

$$W^k(r) = \sum_{i=1}^{N}(C^i(r) \times B_k^i) + b^k(r), \tag{4}$$

where $C^i(r), b^k(r)$ denote the per-pixel coefficient elements and the k_{th} kernel bias at pixel r, B_k^i denotes the low-dimensional basis for the k_{th} kernel.

Sample-Adaptive Element-wise CNN. In low-light conditions, each image is subject to different light distortion model and requires unique non-uniform adjustments. Hence, networks are employed to predict model parameters. Ideally, more adaptive and appropriate models are obtained.

It is critical but difficult to choose the suitable receptive field for a model when facing different contexts and distortions [13]. We design Sample-Adaptive Element-wise CNN to address this issue, as shown in Fig. 2. The degraded image is fed to the CNN, and filtered successively by the multi-layer kernels, which realize the superposition of receptive fields. For each kernel, there is a weighted residual connection so that part of the image data may not be processed by every kernel, thus achieving a variety of receptive fields. Taking a predicted CNN $\mathcal{F}(\cdot)$ with three kernels $\{W^k\}_{k=1}^3$ for low-light image I as an example, the process of filtering can be expressed as:

$$\begin{aligned}\mathcal{F}(I) &= W^3 * [W^2 * (W^1 * I + \alpha^1 I) + \alpha^2 (W^1 * I + \alpha^1 I)] \\ &\quad + \alpha^3 [W^2 * (W^1 * I + \alpha^1 I) + \alpha^2 (W^1 * I + \alpha^1 I)] \\ &\quad - \alpha^3 \alpha^2 \alpha^1 I \\ &= W^3 * W^2 * W^1 * I + \alpha^1 W^3 * W^2 * I \\ &\quad + \alpha^2 W^3 * W^1 * I + \alpha^3 W^2 * W^1 * I \\ &\quad + \alpha^2 \alpha^1 W^3 * I + \alpha^3 \alpha^1 W^2 * I + \alpha^3 \alpha^2 W^1 * I,\end{aligned} \tag{5}$$

where $*$ denotes per-channel convolution and $\{\alpha^k\}_{k=1}^3$ denotes residual connection weights. That means, for a multi-layer filtering network with N kernels, we can integrate filtered results of $2^N - 1$ receptive fields at most, which greatly enriches the diversity of receptive field sizes under fixed kernels prediction costs. By the residual connection weights, the ratio of different receptive fields can be controlled adaptively. By using exponentially increasing kernel size, the proposed method reduces the number of layers in the network and alleviates pixel-level information loss by the deep network, so as to make a better trade-off between receptive field and pixel-level information retention. In addition, we also inherit the method in [8] that obtains a larger kernel using interpolation to further increase the flexibility of the receptive field.

Adaptive fusion. After processing by Sample-Adaptive Element-wise CNNs, we get the reconstruction results from filtered flash and low-light no flash images $\hat{I}_{GT}^f, \hat{I}_{GT}^n$, which are fused guided by importance maps A and A_f for composing features to predict the final enhanced output:

$$\hat{I}_{GT} = A \odot \hat{I}_{GT}^n + A_f \odot \hat{I}_{GT}^f, \tag{6}$$

where $\odot$ denotes the element-wise multiplication.

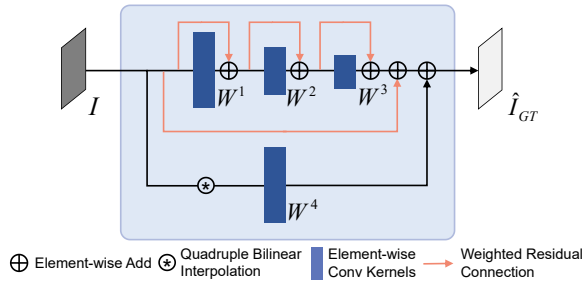

Fig. 2. The design of Sample-Adaptive Element-wise CNN.

2.3. Loss Function

Our models are trained in an end-to-end fashion using back-propagation. We train our model with L2 loss $\mathcal{L}_{\text{rec}}$, gradient loss $\mathcal{L}_{\text{grad}}$ and SSIM loss [14] $\mathcal{L}_{\text{ssim}}$ between the estimated normal-light image and ground truth:

$$\begin{aligned}\mathcal{L}_{\text{rec}} &= ||I_{\text{GT}} - \hat{I}_{\text{GT}}||^2, \\ \mathcal{L}_{\text{grad}} &= \partial_x * |I_{\text{GT}} - \hat{I}_{\text{GT}}| + \partial_y * |I_{\text{GT}} - \hat{I}_{\text{GT}}|, \\ \mathcal{L}_{\text{ssim}} &= 1 - \text{SSIM}(I_{\text{GT}}, \hat{I}_{\text{GT}}),\end{aligned} \tag{7}$$

where ∂_x and ∂_y are horizontal and vertical gradient filters. The loss function of the whole model is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{grad}}\mathcal{L}_{\text{grad}} + \lambda_{\text{ssim}}\mathcal{L}_{\text{ssim}}, \tag{8}$$

where λ_{grad} and λ_{ssim} balance the importance of loss terms.

3. EXPERIMENTAL RESULTS

3.1. Dataset

Our experiments are conducted on the dataset generated by the Flash and Ambient Illuminations Dataset [15]. The image pairs are considered to be noisy and misaligned. We get a low-light clean image by randomly proportionally dimming the normal-light image by a factor of 25 to 50 in linear space and convert it to RGB space. The noise on flash and low-light images is synthesized by Gaussian model, the read and shot noise parameters σ_r and σ_s are sampled uniformly from $[10^{-3}, 10^{-2}]$ and $[10^{-2}, 10^{-1.3}]$. The misalignment between flash and low-light images is simulated by warping flash images with a random homography. In this way, we generate 2,775 pairs of images with different scenes, 100 of which are used for testing and the others for training.

3.2. Implementation Details

In the experiments, we use three kernels on each predicted CNN with weighted residual connection, and the size from front to back are (15×15), (7×7) and (3×3). The size of kernel for filtering the bilinear interpolation images is (15×15). In the training process, we randomly crop the input images into 224×224 patch pairs. Hyperparameters in the loss function are set to $\lambda_{\text{ssim}} = 0.1, \lambda_{\text{grad}} = 1$. Our network is trained

(a) Low-Light (b) Flash (c) DSN (d) FRL (e) DJF (f) CU-Net (g) LSD_2 (h) deepFnF (i) Ours (j) GT

Fig. 3. Visual comparisons with state-of-the-art methods. Note that in the first row there is unnatural illumination in flash images, *e.g.* overexposure or underexposure, and it is resolved in our results. Misalignment as shown in the second row does not disturb our method as well.

Table 1. Quantitative results with state-of-the-art methods.

Methods	DSN [3]	FRL [5]	DJF [6]	CU-Net [4]	LSD_2 [16]	deepFnF [8]	Ours
PSNR	19.14	18.07	20.94	21.95	22.34	24.09	**25.10**
SSIM	0.5266	0.5931	0.4321	0.6007	0.6058	0.6768	**0.7050**
LPIPS	0.5530	0.4909	0.5072	0.5442	0.4095	0.3163	**0.2834**

using the Adam optimizer [17] with a beginning learning rate of 10^{-4}, and decayed to 10^{-5} after 1.2×10^6 iterations, where we totally train the network for 1.5×10^6 iterations.

3.3. Comparison Results

We quantitatively evaluate our proposed method and compare it with the state-of-the-art low-light enhancement method and multi-modal image restoration methods, including Deep Symmetric Network (DSN) [3], CU-Net [4], LSD_2 [16], DJF [6], FRL [5] and deepFnF [8]. PSNR, SSIM and LPIPS [18] are adopted to evaluate the low-light enhancement performance. LPIPS adopts the pre-trained VGG network [19] as the backbone to obtain features from images. We show the scores obtained by different methods in Table 1, demonstrating that our method achieves the best performance.

Visual results are provided in Fig. 3. As shown, without the auxiliary information of flash images, the single input low-light enhancement method DSN [3] is difficult to reconstruct fine structures. And the results obtained by state-of-the-art multi-modal image restoration methods still have obvious degradation such as residual noise (CU-Net [4], DJF [6]), illumination distortion (FRL [5]) and over-smoothing (LSD_2 [16], deepFnF [8]). Comparatively, our method makes more accurate and visual-pleasing reconstruction of details and color. More results can be found on our website[1].

3.4. Ablation Studies

We conduct extensive ablation studies for our proposed network architecture. By replacing the adaptive image fusion module with directly element-wise addition, the model fuses information directly (without adaptive fusion). By replacing the combination way of multiple kernels from our SAECNN to the way of [20], the low-light and flash images can only be filtered by a single filtering kernel (without the design of SAECNN). By replacing the flash input with low-light input, the model fails to get information from the flash image (without flash compensation). We do the above substitutions in turn and observe a significant performance drop as Table 2 shows, though the model sizes are kept almost the same. Hence, all of the components in our dynamic filtering network contribute to performance improvement.

Table 2. Quantitative results of the ablation studies.

baseline	w/ flash	w/ SAECNN	w/ fusion	PSNR	SSIM	LPIPS
✓	✓	✓	✓	**25.10**	**0.7050**	**0.2834**
✓	✓	✓		24.72	0.7016	0.2897
✓	✓			24.29	0.7005	0.2990
✓				23.83	0.6478	0.3581

4. CONCLUSION

In this work, a novel Flash Compensated Dynamic Filtering Network is proposed, introducing flash images as detail structure compensation. The layer-wise regression strategy, sample-adaptive mechanism and coarse-to-fine architecture lead to more efficient utilization of complementary information from the flash and low-light images, making the final output excellent both in illumination and details. Experimental evaluation shows the superiority of our proposed Flash Compensated Dynamic Filtering Network.

[1] https://ellisonkuang.github.io/SAECNN

5. REFERENCES

[1] Stephen M Pizer, R. Eugene Johnston, James P. Ericksen, Bonnie C. Yankaskas, and Keith E. Muller, "Contrast-limited adaptive histogram equalization: Speed and effectiveness," in *Proc. Conf. Vision in Biomedical Computing*, 1990.

[2] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu, "Deep retinex decomposition for low-light enhancement," in *Proc. British Machine Vision Conf.*, 2018.

[3] Lin Zhao, Shaoping Lu, Tao Chen, Zhenglu Yang, and Ariel Shamir, "Deep symmetric network for underexposed image enhancement with recurrent attentional learning," in *Proc. IEEE/CVF Int'l Conf. Computer Vision*, 2021.

[4] Xin Deng and Pier Luigi Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3333–3348, 2020.

[5] Xiongwei Liu, Zehua Sheng, and Huiliang Shen, "Frequency-relevant residual learning for multi-modal image denoising," in *Proc. IEEE Int'l Conf. Image Porcessing*, 2022.

[6] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang, "Joint image filtering with deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell."*, vol. 41, no. 8, pp. 1909–1923, 2019.

[7] Beomjun Kim, Jean Ponce, and Bumsub Ham, "Deformable kernel networks for joint image filtering," *Int'l Journal of Computer Vision*, vol. 129, no. 2, pp. 579–600, 2021.

[8] Zhihao Xia, Michaël Gharbi, Federico Perazzi, Kalyan Sunkavalli, and Ayan Chakrabarti, "Deep denoising of flash and no-flash pairs for photography in low-light environments," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021.

[9] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool, "Dynamic filter networks," in *Proc. Advances in Neural Information Porcessing Systems*, 2016.

[10] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama, "Digital photography with flash and no-flash image pairs," *ACM Trans. Graphics*, vol. 23, no. 3, pp. 664–672, 2004.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Medical Image Computing and Computer-Assisted Intervention Int'l Conf.*, 2015.

[12] Zhihao Xia, Federico Perazzi, Michael Gharbi, Kalyan Sunkavalli, and Ayan Chakrabarti, "Basis prediction networks for effective burst denoising with large kernels," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020.

[13] Talmaj Marinč, Vignesh Srinivasan, Serhan Gül, Cornelius Hellge, and Wojciech Samek, "Multi-kernel prediction networks for denoising of burst images," in *Proc. IEEE Int Conf. Image Porcessing*, 2019.

[14] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncell, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Porcessing*, vol. 15, no. 2, pp. 430–444, 2004.

[15] Yagiz Aksoy, Changil Kim, Petr Kellnhofer, Sylvain Paris, Mohamed Elgharib, Marc Pollefeys, and Wojciech Matusik, "A dataset of flash and ambient illumination pairs from the crowd," in *Proc. European Conf. Computer Vision*, 2018.

[16] Janne Mustaniemi, Juho Kannala, Jiri Matas, Simo Särkkä, and Janne Heikkilä, "LSD_2 – Joint denoising and deblurring of short and long exposure images with cnns," in *Proc. British Machine Vision Conf.*, 2020.

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

[18] Richard Zhang, Phillip Isola, Alexei Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2018.

[19] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int'l Conf. Learn. Representations*, 2015.

[20] Wooyeong Cho, Sanghyeok Son, and Dae-Shik Kim, "Weighted multi-kernel prediction network for burst image super-resolution," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021.