Low-Light RAW Image Enhancement with Additive Parameterization and State Space Model

Shugo Yamashita and Masaaki Ikehara
Keio University, Japan
E-mail: {yamashita, ikehara}@tkhm.elec.keio.ac.jp Tel: +81-45-566-1530

***Abstract*—Low-light image enhancement is a critical yet challenging task in computer vision. Processing RAW sensor data offers significant advantages for this task compared to sRGB images, owing to its higher bit depth and linearity. Leveraging these characteristics, various single-stage and multi-stage architectures have been developed. A recent multi-stage approach decouples the task into RAW denoising and color restoration, with parameter-shared encoders to improve performance and efficiency. However, this parameter-sharing strategy creates a task conflict, as the objectives of these two sub-tasks are not perfectly aligned. Specifically, the encoder for color restoration is required to aggregate task-specific features, not just reconstruct the signal. To address this, we propose a novel network architecture designed for low-light image enhancement. We introduce an Additively-Parameterized Encoder (APEncoder), a refined parameter-sharing mechanism that mitigates the task conflict by enabling a sub-network to learn a set of task-specific parameters that are additively combined with the shared ones. We also present an Attentive State Space Decoder (ASDecoder), which employs a hierarchical structure to efficiently capture global, scene-level characteristics across multiple resolutions. A Mixed-Scale Gated Fusion Module (MSGFM) is developed to improve feature propagation between network stages using multi-scale convolutions. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches, both quantitatively and qualitatively.**

I. Introduction

Developing effective methods for low-light image enhancement is a significant challenge, as images captured under these conditions suffer from poor visibility and high noise. Low-light image enhancement approaches are primarily distinguished by their input data: standard RGB (sRGB) images [1]–[3] or RAW sensor data [4]–[8]. RAW images provide a better foundation for low-light enhancement than sRGB images, due to two fundamental characteristics. First, RAW images contain richer pixel information due to their higher bit depth, which is typically 12 or 14-bit compared to 8-bit for sRGB. Second, as the direct output from the image sensor, RAW images exhibit a linear response to light. In contrast, sRGB images have been processed by a non-linear Image Signal Processor (ISP) pipeline, which complicates the restoration of color and detail from underexposed regions.

Leveraging these inherent benefits of RAW data, a variety of low-light image enhancement approaches [4]–[8] have been researched. SID [4] introduced a large-scale paired dataset for RAW-based low-light image enhancement and proposed a single-stage convolutional network. EEMEFN [5] used a two-stage method that enhances extremely low-light images by combining multi-exposure fusion for color restoration with an edge enhancement module for detail reconstruction. DNF [8] decoupled the enhancement process into denoising in the RAW domain and color restoration in the sRGB domain. In addition, DNF leveraged a parameter-shared encoder to perform both tasks. In this architecture, the encoder switches its role by activating or deactivating local residual shortcuts. While this dual-role design is efficient, it introduces a fundamental challenge: the potentially competing objectives of signal reconstruction for denoising and the aggregation of features specific to the color restoration task. Forcing a single encoder to learn a shared representation that serves both of these not-perfectly-complementary roles can create interference, which may hinder the model's overall learning process.

To overcome this limitation, we introduce an Additively-Parameterized Encoder (APEncoder), a novel approach to parameter sharing. Instead of forcing a single set of weights to serve both tasks, our method defines the weights for the color restoration encoder as the sum of the base noise-removal weights and a task-specific weight set. This allows the color restoration encoder to build upon the foundational features learned for noise removal while also learning specialized features tailored to its own objective. By applying these delta weights specifically to the spatial feature extraction layers, APEncoder mitigates the issue of gradient interference, enabling more effective and stable training for improved color restoration performance.

Furthermore, to address the need for global processing in RAW-to-sRGB conversion, we introduce an Attentive State Space Decoder (ASDecoder). It employs a hierarchical configuration of Attentive State Space Blocks [9] at multiple resolutions, unlike the original MambaIRv2 [9], which is a single-resolution model. This design enables the effective capture of scene-level characteristics.

Finally, to effectively propagate features between network stages, we introduce a Mixed-Scale Gated Fusion Module (MSGFM). This module enhances the Gated Fusion Module (GFM) from DNF [8] by replacing its single-scale convolution with parallel convolutions of multiple kernel sizes. This multi-scale design allows our model to capture a richer set of features, which is crucial for robust color restoration.

Experiments are conducted to evaluate the proposed method on the See-in-the-Dark (SID) [4] dataset, a RAW-based low-

979-8-3315-7206-8/25/$31.00

light image enhancement dataset. Our method achieves quantitatively and qualitatively superior performance compared with state-of-the-art methods. Ablation studies demonstrate that the proposed improvements are effective in RAW-based low-light image enhancement.

II. Proposed Method

A. Overall Architecture

The overall framework of the proposed method is illustrated in Fig. 1. Our proposed network, which builds upon DNF [8], employs a two-stage U-Net architecture for RAW denoising (Stage 1) and color restoration (Stage 2). Given a low-light RAW image I_{RAW}, it is multiplied by the pre-defined amplification ratio [4]. Then, two convolutions and a GELU activation function [10] are applied to expand the channel dimension to C. This shallow feature I'_{RAW} serves as the common input for both stage encoders.

In Stage 1, I'_{RAW} is fed into the Encoder E_1 and Decoder D_1 to generate the denoised RAW image $\hat{I}_{RAW}$. E_1 and D_1 are constructed using the Channel Independent Denoising (CID) block [8], which is a convolutional block.

In Stage 2, the Additively-Parameterized Encoder E_2 (section II-B) processes the shallow feature I'_{RAW}. At each resolution level, its features are fused with those from the corresponding D_1 features via our Multi-Scale Guided Fusion Module (section II-D). Finally, applying the Attentive State Space Decoder (section II-C), we obtain the enhanced sRGB image $\hat{I}_{RGB}$.

We define the loss function $\mathcal{L}$ as a weighted sum of the L1 losses in both RAW and sRGB domains:

$$\mathcal{L} = \|\hat{I}_{RAW} - G_{RAW}\|_1 + \lambda \|\hat{I}_{RGB} - G_{RGB}\|_1, \quad (1)$$

where G_{RAW} and G_{RGB} are the ground truths in the RAW and sRGB domains, and λ denotes a hyperparameter that balances the two terms.

B. Additively-Parameterized Encoder (APEncoder)

In DNF [8], a parameter-shared encoder performs two distinct but complementary tasks: noise removal and signal reconstruction for color restoration. For noise removal, local residual shortcuts are deactivated, allowing the encoder to focus solely on noise estimation. In contrast, for color restoration, these shortcuts are activated to reconstruct the signal by canceling out noise along the shortcut paths.

However, these two roles are not perfectly complementary. The color restoration encoder should not only reconstruct the signal but also aggregate features specific to its task. A potential issue is that the losses from the two tasks can interfere with each other during gradient updates, leading to learning stagnation.

To address this, we propose a refined parameter-sharing method called Additive Parameterization, and employ it in our Additively-Parameterized Encoder (APEncoder). We define the weights of the noise removal encoder E_1 as W_{E1} and the color restoration encoder E_2 as $W_{E1} + \Delta W_{E2}$. This design allows E_2 to share the base parameters W_{E1} with E_1 while incorporating an additional, task-specific parameter set, ΔW_{E2}. As a result, E_2 can learn more effective features for color restoration.

Specifically, following the design of CID, the APEncoder consists of a 7×7 depth-wise convolution (DConv7) and a Multi-Layer Perceptron (MLP). Since DConv7 is primarily responsible for extracting spatial features, the additional parameters ΔW_{E2} in E_2 are applied only to the DConv7. The process of the APEncoder can be summarized as:

$$\begin{aligned} \hat{F} &= \mathrm{DConv7}(F_{in}; W_{E1} + \Delta W_{E2}), \\ F_{out} &= \mathrm{MLP}(\hat{F}; W_{E1}) + F_{in}, \end{aligned} \quad (2)$$

where F_{in} and F_{out} are the input and output features of the APEncoder block. MLP is composed of two point-wise convolutions and a GELU activation function [10].

C. Attentive State Space Decoder (ASDecoder)

In RAW-to-sRGB color conversion, global processing is essential to accurately capture and correct scene-level characteristics such as lighting, color space, and noise. To address this, we introduce the Attentive State Space Decoder (ASDecoder), which efficiently handles global features through a hierarchical configuration of Attentive State Space Blocks (ASBlocks) [9] at multiple resolutions. This hierarchical approach distinguishes our work from MambaIRv2 [9], which applies ASBlocks at a single resolution for super-resolution and denoising tasks.

The core component, the ASBlock, combines an Attentive State Space Module (ASM) for global processing with a Convolutional Feed-Forward Network (ConvFFN) for local processing. The ASM models long-range dependencies with linear computational complexity using the Attentive State Expansion (ASE) mechanism. We modified the original ASE implementation to simplify the embedding calculation. This change allows each module to maintain and optimize its own independent embeddings. In addition, the ConvFFN extracts local features using two point-wise convolutions and a depth-wise convolution.

D. Mixed-Scale Gated Fusion Module (MSGFM)

We introduce the Mixed-Scale Gated Fusion Module (MSGFM) to propagate features from the Stage 1 decoder to the Stage 2 encoder. The MSGFM extends the Gated Fusion Module (GFM) from DNF [8] by incorporating a multi-scale mechanism. While DNF uses only a single 3×3 convolution, which limits feature propagation to a fixed scale, our MSGFM employs convolutions with multiple kernel sizes. This multi-scale design, successful in other image restoration tasks [11], [12], captures a wide range of features that are vital for robust color restoration.

The detailed structure of our MSGFM is presented in Fig. 2. At the l-th ($l \in \{1, 2, \ldots, L\}$) stage, the Stage 1 decoder feature F^l_{D1} and Stage 2 encoder feature F^l_{E2} are concatenated and fed into a point-wise convolution. This result is processed by two parallel depth-wise convolution branches (3×3 and

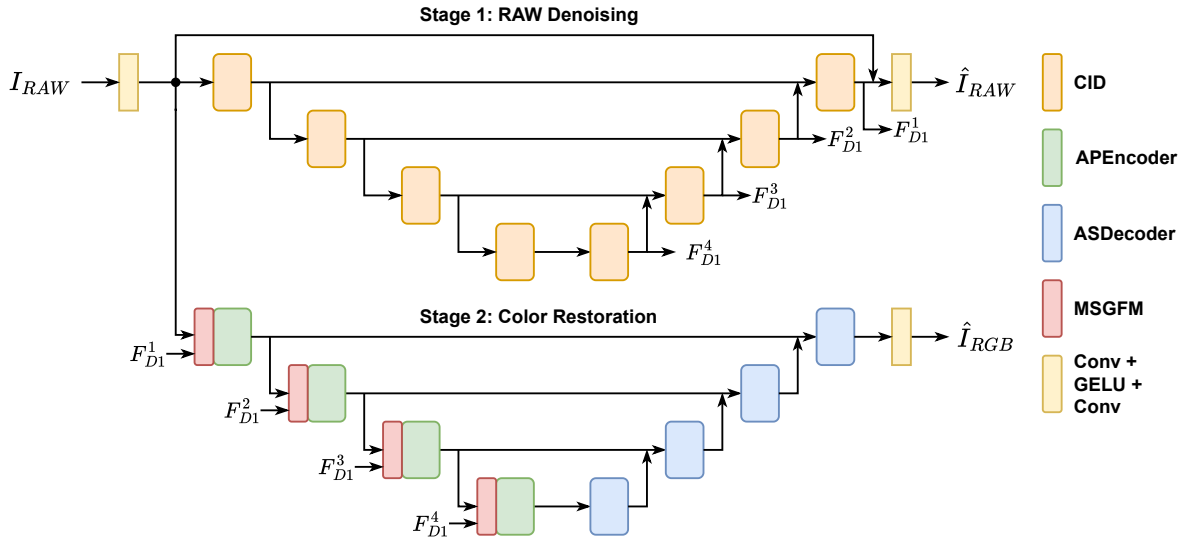

Fig. 1. The overall architecture of our method.

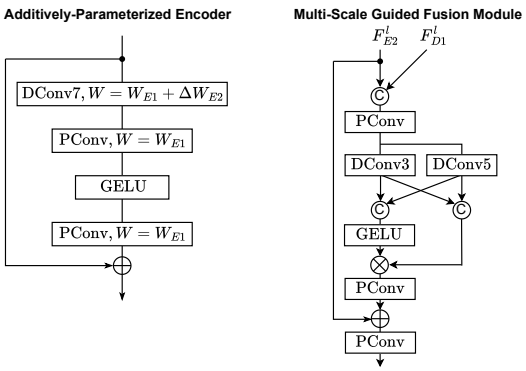

Fig. 2. The architecture of Additively-Parameterized Encoder (APEncoder) and Mixed-Scale Gated Fusion Module (MSGFM).

5×5). The gating is achieved by splitting the channels into two and performing an element-wise product. Subsequently, the features are fused and refined by two point-wise convolutions to yield F^l_{fuse}. Overall, the MSGFM process is defined as:

$$\begin{aligned}
&\hat{F}_a, \hat{F}_b = \text{Chunk}(\text{PConv}([F^l_{E2}, F^l_{D1}])),\\
&F^{con}_{3\times3}, F^{gate}_{3\times3} = \text{Chunk}(\text{DConv3}(\hat{F}_a)),\\
&F^{con}_{5\times5}, F^{gate}_{5\times5} = \text{Chunk}(\text{DConv5}(\hat{F}_b)),\\
&F_{con} = [F^{con}_{3\times3}, F^{con}_{5\times5}], F_{gate} = \text{GELU}([F^{gate}_{3\times3}, F^{gate}_{5\times5}]),\\
&F^l_{fuse} = \text{PConv}(\text{PConv}(F_{con} \odot F_{gate}) + F^l_{E2}),
\end{aligned} \tag{3}$$

where PConv denotes 1×1 convolution, DConv3 and DConv5 represent 3×3 and 5×5 depth-wise convolutions, Chunk splits a tensor into two along the channel dimension, $[\cdot]$ indicates the channel-wise concatenation, and $\odot$ is the Hadamard product.

III. Experiment

A. Experimental Settings

1) Datasets: To compare our proposed method with existing methods, we conduct experiments using the See-in-the-Dark (SID) [4] dataset, a RAW-based low-light image enhancement dataset. It contains 5094 low-light RAW images and their corresponding normal-light ground-truth images. The low-light images are captured with short exposure times of 1/10, 1/25, or 1/30 seconds. Ground-truth images are captured with long exposure times of 10 or 30 seconds, resulting in clean and noise-free images. The SID dataset is divided into a Sony subset and a Fuji subset, which contain images captured by Sony α7S II with a Bayer sensor and a Fujifilm X-T2 with an X-Trans sensor, respectively. For supervised learning, we use the low-light RAW images as input and the ground-truth RAW and sRGB images as supervision targets.

2) Implementation Details: The optimization of our two-stage model is performed in a single training procedure that lasts for 500 epochs. Adam optimizer [19] is used with $\beta_1 = 0.9, \beta_2 = 0.999$. The initial learning rate 2×10^{-4} is gradually decayed using the cosine annealing schedule [20].

The U-Nets in both stages have a depth of four ($l = 4$), with a base embedding dimension C of 32. In the encoder, the number of feature channels is doubled at each level, while the spatial resolution is halved. The decoder symmetrically reverses this process. Downsampling is performed using 2×2 convolutions with a stride of 2, whereas upsampling is implemented with transposed convolutions. The hyperparameter λ in the loss functions is set to 1.

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

1) Compared Methods: We compare our method with state-of-the-art RAW-based low-light image enhancement methods,

TABLE I
QUANTITATIVE COMPARISON ON THE SONY AND FUJI SUBSETS OF SID [4] DATASET. BEST PERFORMANCE IS MARKED IN **BOLD**.

Architecture	Method	Params. ↓	Sony		Fuji	
			PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Single-Stage	SID [4]	7.7 M	28.96	0.787	26.66	0.709
	DID [13]	2.5 M	29.16	0.785	-	-
	SGN [14]	19.2 M	29.28	0.790	27.41	0.720
	LLPackNet [15]	1.2 M	27.83	0.755	-	-
	RRT [16]	0.8 M	28.66	0.790	26.94	0.712
Multi-Stage	EEMEFN [5]	40.7 M	29.60	0.795	27.38	0.723
	LDC [17]	8.6 M	29.56	**0.799**	27.18	0.703
	MCR [6]	15.0 M	29.65	0.797	-	-
	RRENet [7]	15.5 M	29.17	0.792	27.29	0.720
	DNF [8]	2.8 M	30.62	0.797	28.71	**0.726**
	Ours	3.3 M	**30.67**	**0.799**	**28.82**	**0.726**

Input SID [4] MCR [6] DNF [8] Ours GT

Fig. 3. Visual comparison on the images of SID [4] dataset.

TABLE II
ABLATION STUDY OF THE ENCODER PARAMETER SHARING STRATEGY.

Method	Parameter of Encoder		Params. ↓	PSNR ↑	SSIM ↑
	Stage 1	Stage 2			
No sharing	W_{E1}	W_{E2}	3.8 M	30.55	**0.799**
Full Sharing [8]	W_{E1}	W_{E1}	3.3 M	30.54	0.798
Additive Parameterization (Ours)	W_{E1}	$W_{E1} + \Delta W_{E2}$	3.3 M	**30.67**	**0.799**

TABLE III
ABLATION STUDY OF THE STAGE 2 DECODER ARCHITECTURE.

Method	PSNR ↑	SSIM↑
Convolution	29.50	0.793
Transposed Self-Attention [18]	30.53	0.798
ASDecoder (Ours)	**30.67**	**0.799**

TABLE IV
ABLATION STUDY OF THE MIXED-SCALE GATED FUSION MODULE (MSGFM).

Method	PSNR ↑	SSIM↑
GFM [8]	30.57	0.797
MSGFM (Ours)	**30.67**	**0.799**

including single-stage methods [4], [13]–[16] and multi-stage methods [5]–[8], [17].

2) Quantitative Comparison: Tab. I shows that our method outperforms all other competing methods, achieving the highest PSNR/SSIM scores on both the Sony and Fuji subsets. It surpasses the previous best method, DNF [8], with performance gains of 0.05 dB and 0.11 dB, respectively. Notably, this superior performance is achieved with a highly efficient model. Among high-performance multi-stage methods [5]–[8], [17], our model has the second-smallest parameter count, only 0.5M larger than DNF [8], the most compact model.

3) Qualitative Comparison: Fig. 3 shows visual results on the SID [4] dataset. Compared with existing methods [4], [6], [8], our proposed network demonstrates superior performance in noise reduction and the fidelity of color and texture reproduction. This is particularly notable on the lips in the top row and on the umbrella in the bottom row.

C. Ablation Studies

In ablation studies, all models are trained on the Sony subset of SID [4] with the same conditions.

1) Additively-Parameterized Encoder (APEncoder): Tab. II shows that our proposed Additive Parameterization strategy for the encoders is highly effective. The Additively-Parameterized Encoder (APEncoder), which computes the Stage 2 parameters by additively combining stage-specific parameters with those from Stage 1, surpasses simpler strategies like not sharing parameters or using identical parameters for both stages. This confirms that additively updating shared parameters is a beneficial design.

2) Attentive State Space Decoder (ASDecoder): As presented in Tab. III, our proposed ASDecoder is the most effective architecture for the stage 2 decoder. It achieves higher performance than variants using either a convolution block, similar to the Stage 1 decoder, or a transposed self-attention [18].

3) Mixed-Scale Gated Fusion Module (MSGFM): Tab. IV compares our proposed Mixed-Scale Gated Fusion Module (MSGFM) with the Gated Fusion Module (GFM) [8]. By using depth-wise convolutions with both 3×3 and 5×5 kernels, MSGFM outperforms GFM, which relies on a single 3×3 kernel. This result demonstrates the effectiveness of fusing multi-scale features.

IV. Conclusions

In this paper, we presented a novel network architecture specifically designed to address key challenges in RAW-based low-light image enhancement. Our proposed solution integrates three synergistic components. The Additively-Parameterized Encoder (APEncoder) successfully mitigates task interference by allowing for specialized feature learning on top of a shared parameter base. In addition, the Attentive State Space Decoder (ASDecoder) leverages a hierarchical design to efficiently capture long-range, scene-level dependencies essential for accurate color and lighting restoration. Finally, the Mixed-Scale Gated Fusion Module (MSGFM) enhances inter-stage communication by fusing features across different scales. Experimental results demonstrate that our method is more effective than state-of-the-art methods, achieving higher quantitative scores in PSNR and SSIM while also producing visually superior images.

References

[1] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 8, pp. 4225–4238, 2021.

[2] C. Li, C. Guo, L. Han, *et al.*, "Low-light image and video enhancement using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 9396–9416, 2021.

[3] H. Zhou, W. Dong, X. Liu, *et al.*, "Glare: Low light image enhancement via generative latent feature based codebook retrieval," in *European Conference on Computer Vision*, 2024, pp. 36–54.

[4] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Computer Vision and Pattern Recognition*, 2018.

[5] M. Zhu, P. Pan, W. Chen, and Y. Yang, "Eemefn: Low-light image enhancement via edge-enhanced multi-exposure fusion network," in *AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 13 106–13 113.

[6] X. Dong, W. Xu, Z. Miao, *et al.*, "Abandoning the bayer-filter to see in the dark," in *Computer Vision and Pattern Recognition*, 2022, pp. 17 431–17 440.

[7] H. Huang, W. Yang, Y. Hu, J. Liu, and L.-Y. Duan, "Towards low light enhancement with raw images," *IEEE Transactions on Image Processing*, vol. 31, pp. 1391–1405, 2022.

[8] X. Jin, L.-H. Han, Z. Li, C.-L. Guo, Z. Chai, and C. Li, "Dnf: Decouple and feedback network for seeing in the dark," in *Computer Vision and Pattern Recognition*, 2023, pp. 18 135–18 144.

[9] H. Guo, Y. Guo, Y. Zha, *et al.*, "Mambairv2: Attentive state space restoration," in *Computer Vision and Pattern Recognition Conference*, 2025, pp. 28 124–28 133.

[10] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[11] X. Chen, H. Li, M. Li, and J. Pan, "Learning a sparse transformer network for effective image deraining," in *Computer Vision and Pattern Recognition*, 2023, pp. 5896–5905.

[12] S. Yamashita and M. Ikehara, "Image deraining with frequency-enhanced state space model," in *Asian Conference on Computer Vision*, 2024, pp. 3655–3671.

[13] P. Maharjan, L. Li, Z. Li, N. Xu, C. Ma, and Y. Li, "Improving extreme low-light image denoising via residual learning," in *international conference on multimedia and expo*, 2019, pp. 916–921.

[14] S. Gu, Y. Li, L. V. Gool, and R. Timofte, "Self-guided network for fast image denoising," in *International Conference on Computer Vision*, 2019, pp. 2511–2520.

[15] M. Lamba, A. Balaji, and K. Mitra, "Towards fast and light-weight restoration of dark images," *British Machine Vision Conference*, 2020.

[16] M. Lamba and K. Mitra, "Restoring extremely dark images in real time," in *Computer Vision and Pattern Recognition*, 2021, pp. 3487–3497.

[17] K. Xu, X. Yang, B. Yin, and R. W. Lau, "Learning to restore low-light images via decomposition-and-enhancement," in *Computer Vision and Pattern Recognition*, 2020.

[18] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Computer Vision and Pattern Recognition*, 2022, pp. 5728–5739.

[19] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[20] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2017.