ELMformer: Efficient Raw Image Restoration with a Locally Multiplicative Transformer

Jiaqi Ma
School of Computer Science, Wuhan University
Wuhan, China
jiaqima@whu.edu.cn

Shengyuan Yan
School of Computer Science, Wuhan University
Wuhan, China
shengyuan_yan@whu.edu.cn

Lefei Zhang†
Wuhan University
Hubei Luojia Laboratory
Wuhan, China
zhanglefei@whu.edu.cn

Guoli Wang
Horizon Robotics
Beijing, China
guoli.wang@horizon.ai

Qian Zhang
Horizon Robotics
Beijing, China
qian01.zhang@horizon.ai

ABSTRACT

In order to get raw images of high quality for downstream Image Signal Process (ISP), in this paper we present an Efficient Locally Multiplicative Transformer called ELMformer for raw image restoration. ELMformer contains two core designs especially for raw images whose primitive attribute is single-channel. The first design is a Bi-directional Fusion Projection (BFP) module, where we consider both the color characteristics of raw images and spatial structure of single-channel. The second one is that we propose a Locally Multiplicative Self-Attention (L-MSA) scheme to effectively deliver information from the local space to relevant parts. ELMformer can efficiently reduce the computational consumption and perform well on raw image restoration tasks. Enhanced by these two core designs, ELMformer achieves the highest performance and keeps the lowest FLOPs on raw denoising and raw de-blurring benchmarks compared with state-of-the-arts. Extensive experiments demonstrate the superiority and generalization ability of ELMformer. On SIDD benchmark, our method has even better denoising performance than ISP-based methods which need huge amount of additional sRGB training images. The codes are release at https://github.com/leonmakise/ELMformer.

CCS CONCEPTS
- Computing methodologies → Reconstruction; Appearance and texture representations; Computational photography.

KEYWORDS
Raw Images, Image Signal Process, Image Restoration, Transformer

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:
Jiaqi Ma, Shengyuan Yan, Lefei Zhang, Guoli Wang, and Qian Zhang. 2022. ELMformer: Efficient Raw Image Restoration with a Locally Multiplicative Transformer. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3503161.3547767

1 INTRODUCTION

Image restoration is a classic low-level vision task in which raw image restoration is a special but important subtask. From the perspective of Image Signal Processing (ISP), raw data refers to unprocessed information from image sensors and is mainly captured as the initial data source. Raw restoration models output high-quality
raw data whose noises and artifacts are removed. Restoration on raw images can be very essential because those final noisy sRGB images are processed by ISP which contains nonlinear transformation that will extort original shot-and-read noise. Non-negligible shot-and-read noises through the ISP pipeline result in the complex noise distribution of raw images, which can significantly obstacle restoration process. Hence, researchers turn to RAW data where noise and artifacts are uncorrelated and less complex [19, 60]. Although recent models discuss the data structure and noise distribution of raw images, they still neglect the local context information hidden in the raw data structure. Hence, the raw image restoration task faces two main challenges. The first is insufficient exploitation of the raw images by the simple ‘packing’ strategy. The second is the balance between performance and computational cost caused by high-resolution raw images.

For the first challenge, researchers focus on pre-processing strategies to avoid damaging structures of raw images. Chen et al. [10] firstly connect raw images with sRGB ones by CNNs, and introduce a ‘packing’ strategy. Through ‘packing’, the single-channel raw image can be divided into four channels as ‘RGGB’ (Red, Green, Green, Blue) in Bayer patterns. Each channel is half-sized and only contains single color information. Hence, the packed raw images are color-independent in every channel and still keep the spatial dependencies around pixels’ neighborhoods. Depending on this simple strategy, many works [33, 40, 58] related to raw images are proposed. However, Liu et al. [35] reveal the error-prone mode when adapting augmentation methods designed for sRGB images to raw images. The errors easily occur especially when converting raw images to sRGB color space with cropping and flipping. Nevertheless, we notice that raw images are pixel-wise aligned in raw restoration tasks. Simple cropping or flipping operations would not harm the Bayer patterns, and the colors of restored sRGB images are consistent with degraded ones after demosaicing. Although the packing strategy has been a common sense in raw image processing, the spatial dependencies are actually weak. Liang et al. [31] utilize Bi-directional Cross-modal Attention in every stage to enhance both color and spatial branches. From our insights, we prefer to leverage the advantages of color structures and spatial dependencies simultaneously for Transformer-based architectures, and we should ensure single-channel raw images to larger channels by projecting them into multiple subspaces. Hence, our ELMformer replace simple projection module with the Bi-directional Fusion Projection (BFP) module at the input stage.

In terms of the second challenge, we try to make use of those limited data by efficient and effective architectures. Zheng et al. [63] propose a deep convolutional dictionary learning method to learn priors for both representation coefficients and dictionaries. Hu et al. [23] build a pseudo 3-D auto-correlation attention block through 1-D convolutions and a light-weight 2-D structure. Chang et al. [8] design a residual spatial-adaptive block for denoising. Most focus on the noise modeling rather than the basic structural characteristics of raw data. With Transformer-based methods uprising, those methods outperform CNN-based methods and show their promising future. Wang et al. [53] propose an U-shaped Transformer-based network for low-level tasks. Liang et al. [32] device a residual architecture by Swin Transformer. Besides, some studies refer to generating adequate raw data pairs from existing sRGB images. They simulate the ISP pipeline into an invertible procedure to convert sRGB images into raw ones, i.e., UPI [5], CycleISP [60] and PseudoISP [7]. However, the balance of performance and computational cost caused by high-resolution raw images is still a problem, especially for Transformer-based methods like Uformer [53] and SwinIR [32]. They only consider the utilization of Transformer blocks (designed for sRGB images) but ignoring the characteristics of raw data structure. Besides, the limited size of shifted windows results in small receptive fields. Simply enlarging window size will increase the computational consumption quadratically. Therefore, an effective strategy should ensure a larger receptive field and acceptable computational cost. As is shown in Fig. 1, with the proposed Lm-Win Transformer blocks, our ELMformer balance the performance and FLOPs well compared with other SOTAs.

To conclude, we utilize the Bi-directional Fusion Projection (BFP) module and Locally multiplicative Window (Lm-Win) Transformer blocks especially for raw image restoration. Both BFP module and Lm-Win Transformer block are suitable for raw data’s characteristics and can guarantee their efficiency and effectiveness simultaneously. Overall, we summarize the contributions of this paper as follows:

- We analysis the weaknesses of ‘packing’ strategy on raw data and utilize a Bi-directional Fusion Projection module to generate the initial projected features. The BFP module fuses color and spatial information and provides efficient structure priors inside Bayer pattern. Besides, the BFP module enlarges the receptive field and reduces the computation cost at the same image patch scale.
- To further construct the dependencies between pixels, we propose an effective Locally Multiplicative SA (L-MSA) module to enhance the neighborhood dependencies in local sub-windows. The L-MSA module gives a specialized consideration of the characteristics of raw image data and is embedded in every Lm-Win Transformer block.
- Both quantitative and qualitative results demonstrate the superiority of our proposed method. Ablation studies prove the effectiveness of each proposed module in our method. The generalization ability is also proved in this paper.

2 RELATED WORK

2.1 Image Restoration Architecture

Image restoration is often referred to images in sRGB color space, while raw images are commonly excluded. When concerning denoising which is a principal subtask in image restoration, [23, 29, 44, 62] corrupt clean sRGB images with AWGN and other synthetic noise to supervise the training stage. [14, 21, 26] employ U-net architecture to conduct image denoising. Others [24, 28, 57] are self-supervised models. And for deblurring which is also essential for restoration, Nah et al. [41] propose a network which uses multi-scale images as input. GAN and RNN are also employed in [27, 48]. Lu et al. [39] device disentangled representation for self-supervised deblurring. Liang et al. [12] designs a non-blind network to restore night blurry images. Besides, Deng et al. [16] leverage a separable-patch architecture collaborating with a multi-scale integration scheme. For raw-based architecture, Xing et al. [58]...
propose an end-to-end Joint Demosaicing and Denoising (JDD) network based on residual channel attention for joint demosaicing and denoising, and super-resolution. Wu et al. [55] advocate a blind JDD problem and invent a novel divide-and-conquer method to tackle with blind reconstruction from noisy raw images. Liu et al. [36] also devise Invertible Blocks to involve the learning of demosaicing. When concerning raw deblurring, Zhang et al. [61] propose a stacked network consisting of several repetitive DMHPN modules to remove blurring from raw images. Liang et al. [31] design a dual-branch network, which deals with color and spatial information in every blocks. Here, we refer to [31, 59] and utilize as the BFP module in the projection stage.

2.2 Raw Image Pairs

Raw restoration aims to recover high-quality raw images from the degraded raw image, while learning-based models need paired data. Hence, the acquisition of raw image pairs and efficient utilization of paired data are essential. It is time-consuming and labor-intensive to get them from the real world. Most algorithms [4, 43] add synthetic noises to clean images to generate pairs. Note that sRGB synthetic noise is not realistic enough and statistical noise like AWGN is too simple. Shot-and-read noise based on learning and other kinds of learned noise are more realistic instead. The others add shot-and-read noise to clean raw images and feed them into the ISP pipeline to learn the noise changing and extorsion. Cy-cloISP [60] injects shot-and-read noise in raw images along with learned device-agnostic transformation rather than statistical synthetic noise. Tim et al. [5] present a network to ‘unprocess’ images by reverting the process of ISP and add shot-and-read noise in the ‘unprocessing’ process to generate realistic raw sensor measurements. Linh et al. [34] add shot-and-read noise to clean raw data and use GAN to generate realistic noisy raw images. Pseudo-ISF [7] jointly learns ISP pipeline and signal-dependent ‘rawRGB’ noise model to synthesize realistic noisy images. Besides, paired raw images for deblurring can be acquired more easily. The blur-sharp image pairs are taken from the same scene using the sharper one as the ground-truth. In [31], the authors propose a raw deblurring dataset named Deblur-Raw. Many successive raw frames (3-5 frames) are randomly picked from raw videos in various scenes under the same camera setting, where the center one is used as the sharp ground-truth and the rest are blurred images. Considering that we already have some well-collected datasets, our goal is to reach good performances both in quality and quantity. Hence, our ELMformer focuses on how to efficiently and effectively utilize limited raw image pairs rather than extending data scale.

2.3 Vision Transformer

Transformer has been a powerful tool in computer vision recently. The global self-attention is found to be not only suitable for NLP but also CV tasks. For basic CV tasks, PVT [50], Swin Transformer [37], Vit[17] and IPT [11] attain SOTA performances. Specially, Uformer [53] and SwinIR [32] rise as cutting-edge methods in vision transformer for low-level vision tasks such as sRGB image denoising and deblurring. Uformer [53] exploits Local-enhanced Window (LeWin) Transformer block to conduct local transformer operation on multi-scale feature maps which can simultaneously capture long-range features from high-level maps and low-level features from the local information provided by window shifting. Liang et al. [32] organize several Swin Transformer layers together with a residual connection to form a Residual Swin Transformer Block to implement deep feature extraction for image restoration. Image synthesis and image editing are also relevant to our restoration tasks. Cao et al. [6] propose image Local Autoregressive Transformer to better facilitate the locally guided image synthesis. Eser et al. [18] combine convolution and Transformer to enhance the locality of high-resolution image synthesis. In this paper, our ELMformer exploits short-range dependencies from raw images, especially the L-MSA can minimize the performance drop with less computational cost.

3 METHOD

3.1 Network Architecture

Here, we apply the classic U-shaped architecture with the BFP module and Lm-Win Transformer blocks. As shown in Fig. 2, the overall structure of the proposed ELMformer is a hierarchical network with skip-connections between the encoder and the decoder. For instance, given a degraded raw data \( I \in \mathbb{R}^{1 \times H \times W} \), ELMformer firstly applies BFP module to extract low-level features \( X_0 \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}} \) and project single channel to diverse subspaces. Then features are fed to \( K \) encoders just like [25, 45]. Each encoder contains two sequential Lm-Win Transformer blocks and one down-sampling layer. In the downsampling layer, we first reshape the flattened features into 2D spatial feature maps, and then downsample maps and double the channels using 4 \( \times \) 4 convolution with
features are passed to one 1 × 1 convolution along with Sigmoid function. For the Spatial route, we remove the packing strategy and use one 3 × 3 convolution to downsample the input to the same scale as the output of packing. After that, we perform element-wise multiplications on both routes, which cross the color and spatial routes. Finally, the color-based and spatial-based features are concatenated in channel dimension for the following steps.

Notice that the original size of raw image is 1 × H × W, we actually project it into high-dimensional subspaces and downsample it to 1/2, which are features in size of C × H/2 × W/2. Hence, we can take those projected features in smaller size for further efficient processing, and get lower computational costs. Unlike simple Input Projection module which only contains 3 × 3 convolutions for feature extraction, we actually compute the attention of those routes and allocate their weights to another. By this cross multiplication direction, two routes can be enhanced by each other, and the final features contain two parts: one is dominant by color information and the other is dominant by spatial structure. The BFP module utilizes two different properties of raw images: the spatial route considers pixels in the same color filter and the color route maintains the spatial structure of original images.

3.2 Bi-directional Fusion Projection

Denote one raw image as I ∈ R^{1×H×W}, though our goal is to restore a better raw image, we still need to demosaic it to get the final sRGB image. Hence, the correlations inside Bayer patterns which are essential should be preserved well. For most of existing works, they all split the one channel raw image into four channels (RGGB) according to the order of CFA. The strategy is called ‘packing’ and results in one tensor which is 4×H/2×W/2 with separate colors. Considering that raw images contain both color (single pixel) and spatial (between neighbour pixels) information, the rough packing strategy will downscale the resolution of images and breakdown the tight spatial order inside neighbor pixels. To avoid this defect and fully utilize the color and spatial information, we refer to [31, 59] and utilize it as the Bi-directional Fusion Projection module in the projection stage.

As is shown in Fig. 3, the module leads to two routes: Color route for keeping the color consistency and Spatial route for maintaining the spatial structure of original raw images. For the Color route, we perform packing on single-channel raw images and feed them into one 3 × 3 convolution to get their color-based features. Then the
3.3 Locally Multiplicative Window Transformer block

Standard vision transformer blocks [17, 49] reshape images into vectors as tokens to compute global self-attention. To avoid the quadratic computation cost of self-attention on larger resolutions, Liu et al. [37] propose a shifted window strategy. Although it brings the vision transformer into reality, many studies [30, 56] reveal that the local dependencies are not captured well by transformer. For low-level vision tasks such as denoising and deblurring, the local context information is essential due to the complementarity between a degraded pixel and its neighborhoods. Wang et al. [53] enhance the ability of capturing locality by introducing the Locally-enhanced Feed-Forward Network (LeFF). It can replace the traditional MLP layer for better performance. Nevertheless, for raw data which is single-channel and high-resolution, the computational cost is still too large. What’s more, though the packing strategy preserves the color information, the layouts of raw data implicit the spatial characteristics unlike images in sRGB space. Hence, it is essential to utilize the structural similarity of raw data in Transformer-based architectures.

To address the above mentioned issues, we propose a Locally multiplicative Window (Lm-Win) Transformer block, as shown in Fig. 4. The proposed Lm-Win Transformer block is effective and efficient by its special design for raw data, which not only captures long-range dependencies from the self-attention in transformer, but also involves the short-range connections into transformer to capture useful local context. The Self-Attention (SA) module reserves the original Window-based Multi-head Self-Attention (W-MSA) and extends one Locally Multiplicative Self-Attention (L-MSA) branch. For the Forward-Feedback Network (FFN) part, as many researches [30, 51] prove that the depth-wise $3 \times 3$ convolution can enhance the locality of transformer, we apply the same LeFF [53]. Specifically, given the features at the $(l-1)$-th block $X_{l-1}$, the computation of a Lm-Win Transformer block is represented as:

$$
X_l^{SA} = \text{W-MSA}(\text{LN}(X_{l-1})) \otimes \text{L-MSA}(\text{LN}(X_{l-1})) + X_{l-1}
$$

$$
X_l = \text{LeFF}(X_l^{SA}) + X_l^{SA}
$$

where $X_l^{SA}$ and $X_l$ are the outputs of the SA and LeFF parts respectively. LN represents the layer normalization [3].

**Revisiting Window-based Multi-head Self-Attention.** Unlike the vanilla Transformer which computes global self-attention within non-overlapping local windows [53]. Given the 2D feature maps $X \in \mathbb{R}^{C \times H \times W}$ with $C$, $H$ and $W$ being the channel, height and width of the maps, we firstly split those feature maps into non-overlapping windows according to the $H$ and $W$ dimensions with the window size of $M \times M$. In this way, we get the flattened and transposed features $X_j \in \mathbb{R}^{M^2 \times C}$ from each window $j$. Suppose we perform self-attention on those flattened features in each window, and the head number is $k$ and the head dimension is $d_k = C/k$, the complexity of the $k$-th head self-attention in the non-overlapping windows should be: $O((M \times M)^2 \times d_k) = O(M^4 \times d_k)$.

![Figure 5: Illustration of costs between W-MSA and L-MSA.](image)

Here, the $k$-th head self-attention in the non-overlapping windows can be defined as:

$$
X = \{X^1, X^2, ..., X^N\}, \quad N = \frac{H \times W}{M^2}
$$

$$
Y^I_k = \text{Attention}(X^I_kW_k^Q, X^I_kW_k^K, X^I_kW_k^V), \quad i = 1, 2, ..., N
$$

where $W_k^Q, W_k^K, W_k^V \in \mathbb{R}^{C \times d_k}$ represent the projection matrices of the queries, keys, and values for the $k$-th head respectively.

**Locally Multiplicative Self-Attention.** Although W-MSA reduces the computational costs, the crude concatenation of nearby features is lack of consideration. For the local partial areas, the network should avoid learning dependencies from noisy pixels and give more attention to beneficial parts. So we propose the Locally Multiplicative Self-Attention (L-MSA) to enhance the color information and give a better explanation of self-attention.

As is shown in Fig. 5, each pixel-wise location represents single color’s feature. For ISP pipeline, neighborhood pixels are used for demosaicing to make up for the absence of the other color information of a pixel. Those adjacent homochromy pixels can also offer more attention to beneficial parts. So we propose the Locally Multiplicative Self-Attention (L-MSA) to enhance the color information in each window area, we compute its self-attention as what we illustrate before.

$$
X = \{X^1_j, X^2_j, ..., X^L_j\}, \quad L = \frac{H \times W}{2^2}
$$

$$
Z^j_k = \text{Attention}(X^j_kp_k^Q, X^j_kp_k^K, X^j_kp_k^V), \quad j = 1, 2, ..., L
$$

where $p_k^Q, p_k^K, p_k^V \in \mathbb{R}^{C \times d_k}$ represent the projection matrices of the queries, keys, and values for the $k$-th head. Besides, $X^j_kM^2(\frac{I-1}{2})+1$, where $I$ is the relative position bias, whose values are taken from $\hat{b} \in \mathbb{R}^{(2M-1) \times (2M-1)}$ with learnable parameters.
Figure 6: Real noise removal results of ELMformer on SIDD benchmark set. Raw images are fed into an official ISP pipeline for better visualization. Details can be seen when zoomed in. Up: noisy images; Down: denoised results.

\[ X^\text{at}(i-1)+2, \ldots, X^\text{at}i \] are divided sub-windows from \( X^i \), where \( i = 1, 2, \ldots, N \).

Note that the computed self-attention reflects the weights of every color channel i.e., Red, Green, Green and Blue (RGBG) in its corresponding position. Here, we view them as short-range weights and multiply them with our W-MSA in every sub-window. For this part, we can represent it as:

\[
\begin{align*}
Z_k^i &= (Z_k^i(i-1)+1, Z_k^i(i-1)+2, \ldots, Z_k^i(i)) \quad (i=1, 2, \ldots, N) \\
y_k^i &= Y_k^i, Z_k^i \\
x_k^i &= (x_k^1, x_k^2, \ldots, x_k^N)
\end{align*}
\]

\( x_k^i \) is the output of the \( k \)-th head. Then the outputs for all heads \( \{1, 2, \ldots, k\} \) are concatenated and then linearly projected to get the final result. In this way, we actually multiply W-MSA with L-MSA separately according to their CFA locations (RGBG).

By this Locally Multiplicative Self-Attention (L-MSA), all feature maps are rearranged according to CFA-based structures, and focus more on the valid pixel-wise information. It makes up for the local connections in W-MSA, so our transformer block can tackle with severe noisy points. And the later experiments prove this conclusion. As is shown in Fig. 5, the computation cost of L-MSA should be \( O(M^2 \times d_k \times 4) \). More detailed discussions about the computational cost of ELMformer are in Subsection 3.4.

### 3.4 Computational Complexity

To give a clear comparison at the same scale, ELMformer and Uformer [53] are designed with same numbers of blocks and projection features. Here, we theoretically analyze the computational complexity of the Lm-Win Transformer Block of ELMformer and the Le-Win Transformer Block of Uformer. For a batch size of single raw image \( 1 \times H \times W \), ELMformer firstly feeds it into the BFP module and gets one \( C \times \frac{H}{4} \times \frac{W}{4} \) tensor, while Uformer results in one \( C \times H \times W \) tensor. In the following Transformer blocks, our method only processes half-size tensors with the same channel number compared with Uformer. In regard to a single Transformer block under the same window size \( M \), Lm-Win processes \( \frac{HW}{M^2} \) windows with each costing \( O(M^4 \times d_k + M^2 \times d_k \times 4) \), and Le-Win processes \( \frac{HW}{M^2} \) windows with each costing \( O(M^4 \times d_k) \). Hence, for a same raw image, Lm-Win costs \( \frac{HW}{M^2} \times d_k \) while Le-Win needs \( O(\frac{HW}{M^2} \times d_k) \) for one raw image. Considering that both of them set 8 as the window size, which is far smaller than normal size of unprocessed raw images (thousands \times thousands in resolution), ELMformer costs nearly 4× less computational complexity compared with Uformer. As is shown in Fig. 1, when we account into all factors of computation, ELMformer still remains at least one third FLOPs of Uformer, which means that ELMformer reaches higher results with less FLOPs.

### 3.5 Comparisons with Uformer

Our ELMformer has a similar architecture to Uformer [53] which is a U-net structure with vision transformer blocks. However, the Transformer blocks in ELMformer are different from that of Uformer. The Le-win Transformer block of Uformer consists of W-MSA module that conducts window-based self-attention and a forward-feeding network, while our Lm-Win Transformer block contains a novel L-MSA module designed especially for raw images. L-MSA conducts self-attention within subwindows in the non-overlapping shifted windows to produce self-attention weights of different color channels in RAW images. Then our Lm-Win Transformer block multiplies the output weights of L-MSA with the window-based self-attention outputs. The color channel self-attention weights of L-MSA serve as a reference for W-MSA and enhance the feature extraction ability of the Transformer blocks.

### 4 EXPERIMENTS

#### 4.1 Datasets

**Smartphone Image Denoising Dataset (SIDD):** SIDD dataset \(^1\) [1] consists of 320 training image pairs (medium set) and 1280 image pairs (benchmark) for evaluations. These images are all collected with five specific smartphone cameras. The optical sensors of smartphones are of much smaller size compared with professional cameras, so the smartphone-produced images are more noisy but also have high resolution.

**Darmstadt Noise Dataset (DND):** DND dataset \(^2\) [42] contains 50 noisy-clean image pairs from four consumer level cameras. Since the resolution of each image is very high, the dataset is cropped into patches of \( 512 \times 512 \) and finally yields 1000 patches totally.

**Deblur-RAW Dataset**

Deblur-RAW Dataset \(^3\) [31] contains 10252 RAW image pairs for raw deblurring, which contain blurred and corresponding deblurred ones. The whole dataset is split into 8752 and 1500 pairs, which are training and testing sets.

#### 4.2 Experimental Settings

Following the common training strategy of Transformer [49], we use the AdamW optimizer [38] with momentum terms of (0.9; 0.999) and a weight decay of 0.02. We randomly augment the training samples using the horizontal flipping and rotating the images by 90 degrees, 180 degrees, or 270 degrees. We set the window size to 8 \( \times \) 8 in Lm-Win Transformer blocks. For the convenience of downsampling, we stress that the window size in the bottleneck stage is set to 4 \( \times \) 4. The projection dimension number is set to 32 for the BFP module. The cosine decay strategy is also applied to decrease the learning rate to \( 1 \times 10^{-6} \) with the initial learning rate \( 4 \times 10^{-4} \). The single \( \ell_1 \) loss is chosen as the loss function and we show the ablation study of loss functions in Appendix.

---

\(^1\)https://www.eecs.yorku.ca/ kamel/sidd/dataset.php  
\(^2\)https://noise.visinf.tu-darmstadt.de/downloads/  
\(^3\)https://github.com/bob831009/raw_image_deblurring
Table 1: The average metrics on the SIDD benchmark set. *: results in italic are from the official pretrained model trained by additional data.

| Method       | PSNR  | SSIM  |
|--------------|-------|-------|
|              | $t/r$ | $t/s$ | $t/r$ | $t/s$ |
| EPLL [64]    | 40.73 | 25.19 | 0.9350 | 0.8420 |
| BM3D [15]    | 45.52 | 30.95 | 0.9800 | 0.8630 |
| KSVD [2]     | 43.26 | 27.41 | 0.9690 | 0.8320 |
| WNNM [20]    | 44.85 | 29.54 | 0.9755 | 0.8885 |
| FoE [46]     | 43.13 | 27.18 | 0.9690 | 0.8120 |
| TNRD [13]    | 42.77 | 26.99 | 0.9450 | 0.7440 |
| UPI* [5]     | 51.54 | 38.91 | 0.9920 | 0.9530 |
| UPI [5]      | 42.23 | 28.39 | 0.8880 | 0.6360 |
| CycleISP* [60] | 51.75 | 39.24 | 0.9930 | 0.9550 |
| CycleISP [60] | 47.98 | 35.02 | 0.9500 | 0.8460 |
| ELMformer    | 51.94 | 39.50 | 0.9930 | 0.9570 |

Table 2: The average metrics on the DND dataset. *: results in italic are from the official pretrained model trained by additional data.

| Method       | PSNR  | SSIM  |
|--------------|-------|-------|
|              | $t/r$ | $t/s$ | $t/r$ | $t/s$ |
| EPLL [64]    | 40.73 | 25.19 | 0.9350 | 0.8420 |
| BM3D [15]    | 46.64 | 37.78 | 0.9724 | 0.9308 |
| KSVD [2]     | 45.54 | 36.59 | 0.9676 | 0.9162 |
| WNNM [20]    | 46.30 | 37.56 | 0.9707 | 0.9313 |
| FoE [46]     | 45.78 | 35.99 | 0.9666 | 0.9042 |
| TNRD [13]    | 44.97 | 35.57 | 0.9624 | 0.8913 |
| UPI* [5]     | 48.89 | 40.17 | 0.9824 | 0.9623 |
| UPI [5]      | 48.44 | 39.47 | 0.9802 | 0.9508 |
| CycleISP* [60] | 49.13 | 40.50 | 0.9830 | 0.9655 |
| CycleISP [60] | 48.75 | 39.84 | 0.9812 | 0.9541 |
| ELMformer    | 48.84 | 40.06 | 0.9816 | 0.9560 |

4.3 Compared Methods
For a complete and fair comparison, we collect several denoising methods, which includes traditional blind ones, CNN-based ones and some simulated ISP pipelines. Besides, we collect several SOTA raw deblurring methods.

Of those denoising methods, EPLL [64], BM3D [15], KSVD [2], WNNM [20] and FoE [46] are optimization based methods for denoising. TNRD [13], UPI [5] and CycleISP [60] are CNN-based methods. Besides, UPI [5] and CycleISP [60] also belong to ISP based pipelines which simulate the process of generating sRGB and raw images. Note that UPI [5] and CycleISP [60] need large-scale sRGB images to train the ISP pipeline, so we mark them with * in Tab. 1 and Tab. 2. For other methods without *, we follow the authors’ instructions and train them only with SIDD medium set for a fair comparison.

4.4 Quantitative Results
Here, we first compare our approach with the aforementioned SOTAs on both denoising benchmarks: SIDD and DND, and then evaluate on one deblurring benchmark: Deblur-RAW. Most results are from the official leaderboard on their websites or corresponding papers. Only some missing results are tested with authors’ pretrained models and submitted to the online server for evaluation. For all three benchmarks, we record the PSNR and SSIM on Raw/Raw and Raw/sRGB space. Considering that only SIDD and Deblur-Raw have training sets, we evaluate DND by the model trained on SIDD medium dataset with no more external data.

From Tab. 1, our ELMformer reaches the highest scores on all the metrics. ELMformer reaches 51.94 dB on PSNR and even surpasses UPI* [5] and CycleISP* [60] which are trained on huge additional data. The ELMformer has at least 3.96 dB and 4.48 dB improvements over CycleISP on raw and sRGB color space, respectively. Hence, our ELMformer can outperform other methods on raw denoising from smartphones.

The results illustrated in Tab. 2 differ. ELMformer is superior to most of methods except UPI* and CycleISP*. Our ELMformer only has 0.10 dB and 0.22 dB promotion over CycleISP on raw and sRGB color space, respectively. Notice that UPI* and CycleISP* are trained on other generated large-scale data, while ELMformer only needs several hundred pairs of raw data, we suppose that those
generated raw images can further extend the generalization ability. ELMformer still has a superiority over other methods at the same scale of training data. Besides, from Tab. 1 and Tab. 2, we can find that ELMformer which is designed for raw restoration has an edge on raw color space compared with other methods and this advantage is weaken on sRGB color space.

We also evaluate on the raw deblurring task in Tab. 3. It is obvious that those SOTAs trained on the sRGB color space have a gap over methods designed for raw deblurring. Even DMPHN_raw [61] has minor superiority over DMPHN_rgb [61]. For three raw deblurring methods, ELMformer still exceeds others in all metrics except SSIM in sRGB color space but outperforms RID a lot in PSNR.

### 4.5 Qualitative Results

Then we show some representative results from three datasets for visual comparison. The raw images are fed into existing ISP pipelines and illustrated in sRGB space.

From Fig. 6, our method removes noisy points in SIDD and keeps the details successfully. It indicates that ELMformer can deal with real noises from mobile devices perfectly. Due to the space limit, we also illustrate the results of DND and Deblur-RAW datasets in Appendix. For DND, ELMformer can still effectively remove real noises and maintain the structure of locality even without training set in DND images’ domain. For deblurring, our ELMformer has better visual results in deblurring. It further proves that methods trained on raw images supply color and spatial information and lead to clear fine-grained details.

### 4.6 Ablation Studies

To evaluate the effectiveness of individual parts in our ELMformer, we perform ablation study on denoising tasks and show the average PSNR and SSIM. As is shown in Table. 4, we take W-MSA [53] as basic blocks of our baseline and compare it along with BFP and L-MSA in the same U-shaped architecture.

Comparing the first and second rows, it is confirmed that W-MSA can perform better with more computational cost as we expect. It reflects that although the FLOPs is nearly quartered, BFP module almost make up the drop in performance by fusion of color architecture and spatial information. Then when it comes to the first and third rows, we can find that L-MSA improves performance by enhancing short-range dependencies between local neighbor pixels while adding little computational cost. At last, the fourth row represents our proposed ELMformer. Compared with other three rows, ELMformer keeps a balance between performance and computational cost. It outperforms the baseline in all metrics while keeps only 1/3 FLOPs. By integrating BFP and L-MSA modules, ELMformer achieves the better results and lower cost. More ablation studies are in Appendix.

### 4.7 Generalization Test

Here, we firstly evaluate our denoising model on some extreme conditions to prove its powerful generalization ability. We select SID [10], ELD [54] and HDR+ [22] datasets to separately show the denoising ability in short exposed, extremely dark and high-dynamic range conditions. Then we also add some synthetic noises to show its modeling ability. To be specific, we add AWGN and Uniform noises to the clean raw data of SIDD [1] as the noisy ones, and directly apply our pretrained model of SIDD medium set to denoise.

In Fig. 7, we show the noisy and restored images which are processed by a simple ISP pipeline. Compared with the noisy images, our method can produce noise-free outputs of high quality in several extreme conditions and maintain local details simultaneously. Also, we denote some noisy raw images taken by mobile devices, i.e., OPPO and XiaoMi, and the visualized results are in Appendix.

## 5 CONCLUSIONS

In this paper, we propose an efficient locally multiplicative Transformer called ELMformer for raw restoration. In contrast to existing CNN-based structures, our model build upon the BFP module and Lm-Win Transformer blocks. Two core components: BFP and L-MSA, are designed for bi-directional (color and spatial) feature fusion in the projection stage and efficient self-attention with stronger short-range connections. Extensive experiments demonstrate that ELMformer achieves competitive results on the raw denoising and deblurring tasks with less computation cost. More qualitative results from several datasets reveal that our pretrained model maintains a strong generalization ability.

## ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant 62122060, the Special Fund of Hubei Luojia Laboratory under Grant 220100014, and the Fundamental Research Funds for the Central Universities under Grant 2042021kf0196. The numerical calculations in this work had been supported by the supercomputing system in the Supercomputing Center of Wuhan University.

---

4. https://github.com/AbdoKamel/simple-camera-pipeline
REFERENCES

[1] Abdelrahman Abdelhamied, Stephen Lin, and Michael S. Brown. 2018. A High-Quality Denoising Dataset for Smartphone Cameras. In CVPR. 1892–1700.

[2] Michal Aharon, Michael Elad, and Alfred M. Bruckstein. 2006. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. TSP. 54, 11 2006). 4331–4322.

[3] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. CoRR abs/1607.06450 (2016).

[4] Long Bao, Zengli Yang, Shuangqiang Wang, Dongwong Bai, and Jungwon Lee. 2020. Real Image Denoising Based on Multi-Scale Residual Dense Block and Cascaded U-Net with Block-Connection. In CVPR. 1823–1831.

[5] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. 2019. Unprocessing Images for Learned Raw Denoising. In NeurIPS. 13106–13105.

[6] Chenjie Cao, Yuxin Hong, Xiang Li, Chengrong Wang, Chengming Xu, Yanwei Fu, and Xiangyang Xue. 2021. The Image Local AutoRegressive Transformer. In NeurIPS. 18433–18445.

[7] Yue Cao, Xiaohu Wu, Shuran Qiu, Xiao Liu, Zhongqin Wu, and Wengang Zuo. 2021. Pseudo-ISP: Learning Pseudo-In-camera Signal Processing Pipeline from A Color Image Denoiser. CoRR abs/2103.10234 (2021).

[8] Meng Chang, Qi Li, Huijun Feng, and Zhihua Xu. 2020. Spatial-Adaptive Network for Single Image Denoising. In ECCV, Vol. 13275. 171–187.

[9] Pierre Charbonnier, Laure Blanc-Féraud, Gilles Aubert, and Michel Barlaud. 1994. Two Deterministic Half-Quadratic Regularization Algorithms for Computed Imaging. In ICP. 168–172.

[10] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. 2018. Learning to See in the Dark. In CVPR. 3291–3300.

[11] Hanjiang Chen, Yunchen Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. 2021. Pre-Trained Image Processing Transformer. In NeurIPS. 12299–12310.

[12] Liang Chen, Jiawen Zhang, Jinhuan Pan, Songnan Lin, Faming Fang, and Jimmy S. Ren. 2021. Learning a Non-Blind Denoising Network for Night Blurry Images. In NeurIPS. 15042–15054.

[13] Yunjin Chen, Wei Yu, and Thomas Pock. 2015. On Learning Optimized Reaction Diffusion Processes for Effective Image Restoration. In CVPR. 5261–5269.

[14] Shen Cheng, Yujing Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, and Jue Wang. 2019. Learning Raw Image Denoising With Bayer Pattern Uniformization and Bayer Pre-Serving Augmentation. In CVPR. 2070–2077.

[15] Yang Liu, Zhenyu Qin, Saeed Anwar, Pan Ji, Dongwoo Kim, Sabrina B. Caldwell, and Tom Gedeon. 2021. Invertible Denoising Network: A Light Solution for Real Noise Removal. In CVPR. 13365–13374.

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yuxian Wei, Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In ICCV. 9992–10000.

[17] Ilya Loushchov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In ICLR.

[18] Boyu Lu, Jun-Cheng Chen, and Rama Chellappa. 2019. Unsupervised Domain-Specific Deblurring via Disentangled Representations. In NeurIPS. 10225–10234.

[19] Yucheng Lu and Seung-Won Jung. 2022. Progressive Joint Low-Light Enhancement and Noise Removal for Raw Images. TIP. 31 (2022), 2390–2404.

[20] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. 2017. Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring. In CVPR. 257–265.

[21] Tobias Pfotz and Stefan Roth. 2017. Benchmarking Denoising Algorithms with Real Photographs. In NeurIPS. 2750–2759.

[22] Tobias Pfotz and Stefan Roth. 2018. Neural Nearest Neighbors. In NeurIPS. 1095–1106.

[23] Chao Ren, Xiaohai He, Chuncheng Wang, and Zhibo Zhao. 2021. Adaptive Consistency Prior Based Deep Network for Image Denoising. In CVPR. 8596–8606.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In MICCAI. Vol. 9351. 234–241.

[25] Stefan Roth and Michael J. Black. 2009. Fields of Experts. IJCV 82, 2 (2009), 205–229.

[26] Peter Shaw, Jakub Uzkorecht, and Ashish Vaswani. 2018. Self-Attention with Relative Position Representations. In NAACL-HLT. 464–468.

[27] Xin Tao, Hongyuan Gao, Xiaoyong Shen, Jian Wang, and Jiaya Jia. 2018. Scale-Recursive Network for Deep Image Deblurring. In CVPR. 8174–8182.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakub Uzkorecht, Loton Ljons, Aidan N. Gomez, Lukaus Kaisler, and Illia Polosukhin. 2017. Attention is All You Need. In NeurIPS. 5998–6008.

[29] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ling Piao, and Liao Shao. 2021. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In ICCV. 548–558.

[30] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ling Piao, and Liao Shao. 2021. PVTv2: Improved baselines with pyramid vision transformer. Computational Visual Media 8, 3 (2022), 1–10.

[31] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. TIP 13, 4 (2004), 600–612.

[32] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Haoqiang Li. 2022. Uformer: A General U-Shaped Transformer for Image Restoration. In CVPR. 17683–17693.

[33] Kaizuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. 2020. A Physically-Based Noise Formation Model for Extreme Low-Light Raw Denoising. In CVPR. 2755–2764.

[34] Fangfang Wu, Tao Huang, Weisheng Dong, Guangming Shi, Zhonglong Zheng, and Xi Lin. 2021. Toward Blind Joint Demosaicing and Denoising of Raw Color Image Filter Array Data. Neurocomputing 453 (2021), 369–382.

[35] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. 2021. CVT: Integrating Convolutions to Vision Transformers. In NeurIPS. 212–221.

[36] Xiaohao Wu, Ming Liu, Yue Cao, Dongwei Ren, and Wangming Zuo. 2020. Unpaired Learning of Deep Image Denoising. In ECCV, Vol. 12349. 3526–3528.

[37] Wenzhong Xing and Karen O. Egiazarian. 2021. End-to-End Learning for Joint Image Demosaicing, Denoising and Super-Resolution. In CVPR. 3007–3016.
A ABLATION STUDIES
Here, we further examine the effectiveness of the channel dimension in Tab. 5. C denotes the channel dimension of projection module.

It is noteworthy that the computational cost will increase a lot with the increase of channel number, while the performance promotion is not significant. It reflects that 32 is a trade-off for the BFP module which can bring effective performance gains while maintaining low FLOPs.

| C  | PSNR  | SSIM  | GFLOPs |
|----|-------|-------|--------|
| 16 | 51.84 | 39.37 | 0.993  | 0.956  | 0.91  |
| 32 | 51.94 | 39.50 | 0.993  | 0.957  | 3.55  |
| 48 | 52.00 | 39.56 | 0.993  | 0.957  | 7.91  |
| 64 | 52.03 | 39.60 | 0.993  | 0.957  | 14.01 |

Table 5: Ablation studies of projection dimension on SIDD benchmark.

To choose the optimal loss function for our ELMformer, we perform ablation study on SIDD benchmark. As is shown in Table. 6, we apply $\ell_1$, $\ell_2$ and Charbonnier [9] loss as alternative losses to supervise the training stage, and show the average PSNR and SSIM. Here, Char is short for the Charbonnier loss.

It is clear that $\ell_1$ performs worst among those loss functions. It reflects that the errors in restoration tasks should not be enlarged. Comparing the first and third rows, $\ell_1$ loss shows a slight promotion than Charbonnier loss in PSNR. Hence, we finally choose $\ell_1$ loss to supervise the training stage.

| Losses | PSNR  | SSIM  |
|--------|-------|-------|
|        | r/r   | r/s   | r/r   | r/s   |
| $\ell_1$ | 51.94 | 39.50 | 0.993 | 0.957 |
| $\ell_2$ | 51.64 | 39.05 | 0.993 | 0.954 |
| Char   | 51.93 | 39.47 | 0.993 | 0.957 |

Table 6: Ablation studies of losses on SIDD benchmark.

B GENERALIZATION TEST IN EXTREME CONDITIONS
In this section, we evaluate our denoising model on some extreme conditions to prove its powerful generalization ability. We collect some noisy raw images taken by mobile devices, i.e., Vivo and XiaoMi. Then we apply ELMformer to denoise those images and visualize them.

Here, we also try our model on other raw images captured by mobile devices from other brands, such as Vivo and XiaoMi. Notice that both SIDD and DND do not contain data from Vivo and XiaoMi, and sensors result in different domains of raw images. From Fig. 8, we observe that noisy points can be reduced well even without any training data from this domain.

C MORE VISUAL RESULTS
Here, we illustrate more denoising and deblurring visual results of our methods and SOTAs for comparison. In Fig. 9, those raw images are processed by official ISP pipelines online, and we collect their feedback in RGB images. For Fig. 10, we also apply a simple ISP to raw images. To give an intuitive impression, we try our best to maintain the originality of images with fine-grained parts.

For denoising in Fig. 9, ELMformer can still effectively remove real noises and maintain the structure of locality even without training set in DND images’ domain. It indicates that ELMformer can deal with real noises from mobile devices perfectly.

For deblurring in Fig. 10, ELMformer has better visual results than other methods. It further proves that methods trained on raw images supply color and spatial information and lead to clear fine-grained details.
Figure 9: Real noise removal results on DND dataset. Raw images are fed into an existing ISP pipeline for better visualization.

Figure 10: Deblur results on Deblur-Raw dataset. Raw images are fed into an existing ISP pipeline for better visualization.