Enhancing RAW-to-sRGB with Decoupled Style Structure in Fourier Domain

Xuanhua He[1,2*], Tao Hu[1,2*], Guoli Wang[3], Zejin Wang[3], Run Wang[3], Qian Zhang[3], Keyu Yan[1,2], Ziyi Chen[4], Rui Li[1], Chenjun Xie[1], Jie Zhang[1†], Man Zhou[5†]

[1]Hefei Institutes of Physical Science, Chinese Academy of Sciences
[2]University of Science and Technology of China
[3]Horizon Robotics
[4]Tencent Technology
[5]Nanyang Technological University
{hexuanhua,ht_simon,keyu}@mail.ustc.edu.cn, {cjxie,zhangjie}@iim.ac.cn,manzhountu@gmail.com

Abstract

RAW to sRGB mapping, which aims to convert RAW images from smartphones into RGB form equivalent to that of Digital Single-Lens Reflex (DSLR) cameras, has become an important area of research. However, current methods often ignore the difference between cell phone RAW images and DSLR camera RGB images, a difference that goes beyond the color matrix and extends to spatial structure due to resolution variations. Recent methods directly rebuild color mapping and spatial structure via shared deep representation, limiting optimal performance. Inspired by Image Signal Processing (ISP) pipeline, which distinguishes image restoration and enhancement, we present a novel Neural ISP framework, named FourierISP. This approach breaks the image down into style and structure within the frequency domain, allowing for independent optimization. FourierISP is comprised of three subnetworks: Phase Enhance Subnet for structural refinement, Amplitude Refine Subnet for color learning, and Color Adaptation Subnet for blending them in a smooth manner. This approach sharpens both color and structure, and extensive evaluations across varied datasets confirm that our approach realizes state-of-the-art results. Code will be available at https://github.com/alexhe101/FourierISP.

Introduction

The Image Signal Processing (ISP) pipeline, responsible for transforming RAW data captured by camera sensors into sRGB images, involves a series of low-level vision tasks including demosaicing, denoising, gamma correction, white balance, and color correction (Ramanath et al. 2005). Traditionally, these individual subprocesses are executed via independent algorithms, often requiring significant manual parameter adjustments (Zhou and Glotzbach 2007). With the rise of mobile photography, smartphones have become the preferred choice for image capture due to their portability. However, the inherent limitations of sensor size and aperture in comparison to DSLR cameras pose challenges for mobile devices in achieving DSLR-like image quality. To bridge this gap, learning RAW-to-sRGB mapping through deep ISP models holds great promise (Ignatov, Van Gool, and Timofte 2020; Ignatov et al. 2020). These models hold the potential to convert mobile RAW data into high-quality sRGB images resembling those captured by DSLR cameras, without the need for manual fine-tuning.

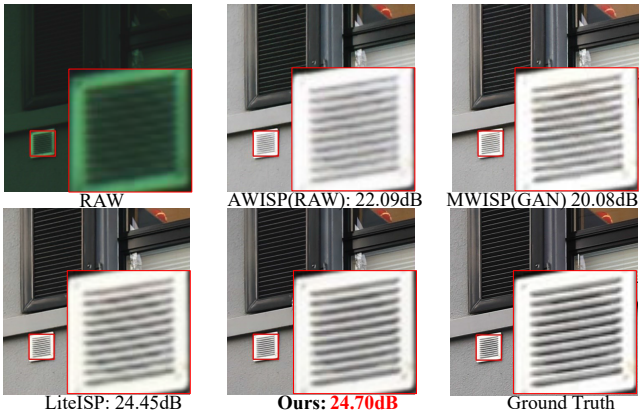

Figure 1: Results from the ZRR dataset. Our approach results in clear textures, surpassing other methods.

Recent deep learning-based ISP methods that utilize mobile phone RAW images have shown promise in producing RGB images comparable to those from DSLR cameras (Dai et al. 2020; Ignatov et al. 2021). This innovation enables cost-effective mobile sensors to deliver visually appealing results. However, existing techniques often focus solely on the Raw-to-RGB conversion as a color mapping task, overlooking the crucial spatial relationship between different image types. This oversight leads to outputs with reduced clarity. In conventional ISP pipelines, processes like denoising, demosaicing, white balance, and gamma correction serve two interconnected yet distinct purposes: image restoration, which preserves spatial structure, and image enhancement, which involves color adjustments (Liang et al. 2021). The effectiveness of enhancement relies heavily on successful

[*]Co-first authors contributed equally. This work was done when Xuanhua He was an intern at Horizon Robotics. [†] Corresponding author.

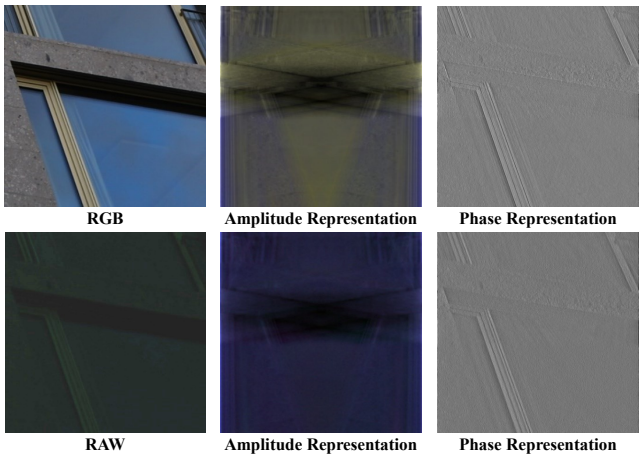

Figure 2: Illustration of image amplitude and phase representation.

restoration; any shortcomings in demosaicing or denoising can hinder the learning of color information. Many widely-used models adopt one-stage design strategies that merge the learning of various ISP subtasks. Unfortunately, this approach limits the representation capability of CNN models and results in images lacking clear details and suffering from color distortions. Therefore, developing an efficient method to separate style and structural information is essential for unlocking enhanced performance within ISP frameworks.

However, achieving effective style-structure decoupling in the spatial domain requires complex loss functions, which can be challenging (Yang et al. 2023). As a result, we shift our focus to the frequency domain, where the powerful prior of Fourier transform offers a promising solution for style and structure decoupling, as depicted in the Figure 2. The Fourier transform of an image yields its amplitude, representing the style, and its phase, representing the structure.

Based on the aforementioned insights, we introduce a novel Neural ISP framework named Fourier-ISP. Unlike previous methods, our approach utilizes the Fourier prior to decouple and optimize color knowledge and structure representation. The network consists of three crucial subnets: the Phase Enhance Subnet (PES) for enhancing spatial structure and fine textures, the Amplitude Refine Subnet (ARS) for learning precise color information, and the Color Adaptation Subnet (CAS) responsible for transmitting color information to the phase-enhanced feature, thereby obtaining rich spatial details and ensuring precise color output. Through extensive evaluations on multiple datasets, our method showcases state-of-the-art results in qualitative and quantitative assessments, while also demonstrating robust transferability.

Our contribution can be summarized as follows: 1) In this work, we present a novel approach utilizing the Fourier prior to decouple style and structure in the Raw-to-RGB mapping process. By separately optimizing the style and spatial structure of RAW images, we achieve highly accurate raw-to-RGB mapping results. 2) We propose the Fourier-ISP framework, consisting of three specialized sub-networks: PES, ARS, and CAS. This well-crafted architecture enables distinct subnets to acquire specific expertise, resulting in superior raw-to-RGB performance. 3) The proposed approach outperforms the state-of-the-art on multiple datasets, as shown in extensive quantitative and qualitative experiments.

Related Work

Deep Learning for ISP

In recent years, the integration of deep learning into ISP pipelines has gained significant attention due to the complexity manual adjustments in traditional pipelines. Some approaches target specific ISP modules, like image denoising (Cheng et al. 2021), demosaicing (Liu et al. 2020), and tone mapping (Hu, Chen, and Allebach 2022), while others aim to overhaul the entire pipeline using neural network models. In prior works, both RGB and RAW images were acquired using the same device, as shown by DeepISP's (Schwartz, Giryes, and Bronstein 2018) end-to-end structure and CameraNet's (Liang et al. 2021) division of the task into restoration and enhancement stages, guided by software-generated ground truth. Among recent endeavors, the Pynet and ZRRdataset (Ignatov, Van Gool, and Timofte 2020) raise the challenging task of mapping mobile RAW images to DSLR camera RGB images, complicated by resolution discrepancies and spatial misalignment due to dual-device capture. Notable contributions include MW-ISPNet (Ignatov et al. 2020) leveraging MWCNN (Liu et al. 2018) for utilizing multi-scale features, AWNet (Dai et al. 2020) employing attention mechanisms for refined color learning and misalignment handling, and LiteISP's (Zhang et al. 2021) lightweight design with optical flow alignment for fine texture output. LWISP (Chen and Ma 2022) adopt distillation for efficiency, achieving a parameter-effectiveness balance. However, prevailing approaches treat RAW-to-RGB mapping as straightforward regression tasks, often ignoring disparities in structural and color information. While CameraNet recognizes these disparities, it requests intermediate result supervision from manual labeling. Our method stands apart by explicitly promoting style and color learning through frequency domain style-structure decoupling.

Fourier Transform in Computer Vision

In the realm of deep learning, the Fourier transform has attracted significant interest due to its distinctive attributes. Leveraging its global properties, FFC (Chi, Jiang, and Mu 2020) devised convolutional modules, effectively utilizing global information processing while maintaining minimal computational overhead. LAMA (Suvorov et al. 2022) employed FFC to construct an UHD image inpainting network by utilizing its global feature. In low-level vision, frequency domain attributes elevate high-frequency image details in super-resolution tasks (Zhou et al. 2022), while image restoration tasks dissociate degradation features using Fourier transform (Zhou et al. 2023). However, its style-structure decoupling attributes have not been fully explored in the low level vision community and its application within the RAW-to-RGB remains uncharted.

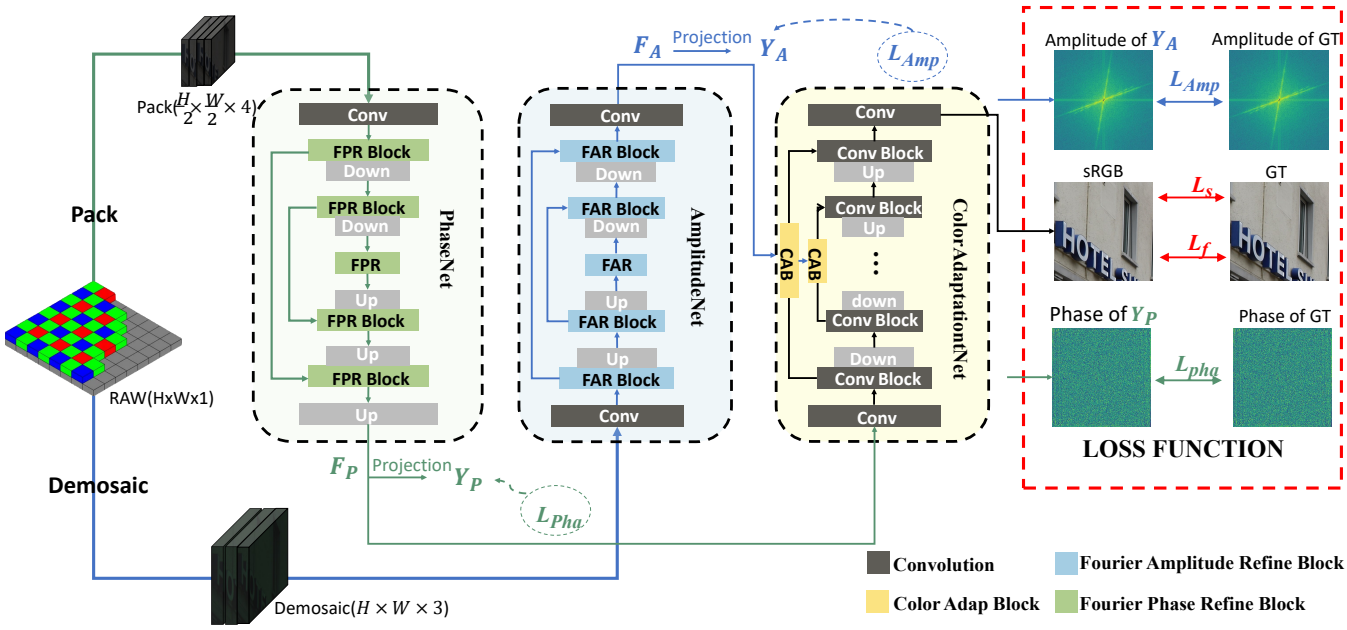

Figure 3: Our Model Framework. We employ separate processing for RAW images through packing and demosaicing. These processed images are subsequently fed into PES and ARS to learn the spatial details and style information of the image, respectively. Finally, we integrate the style information into the spatial features using the CAS and produce the final output.

Method

We leverage Fourier transform to decouple style and structure, enabling separate optimization of these two components. This section begins with an overview of the basic knowledge of Fourier transform, followed by an introduction of the network architecture and the loss functions.

Fourier Transform of Image

The Fourier Transform, widely employed in image processing, enables the conversion of signals into the frequency domain. The Fourier transform of an input image $x \in R^{h\times w}$ can be defined as follows:

$$\mathcal{F}(x)(u,v) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h,w) e^{-j2\pi(\frac{h}{H}u + \frac{w}{W}v)} \tag{1}$$

The amplitude and phase components are described as:

$$\mathcal{A}(x)(u,v) = [R^2(x)(u,v) + I^2(x)(u,v)]^{\frac{1}{2}}, \tag{2}$$

$$\mathcal{P}(x)(u,v) = \arctan[\frac{I(x)(u,v)}{R(x)(u,v)}] \tag{3}$$

where $I(x)$ and $R(x)$ indicate imaginary and real parts of the image's frequency representation $\mathcal{F}(x)$, correspondingly.

The Fourier transform offers two distinctive attributes. Firstly, its global nature arises from the weighted summation of spatial domain values, thus inherently containing global information. By leveraging this, we go beyond the limitations of local receptive fields in convolutional neural networks, acquiring pivotal global context for effective color mapping. Furthermore, the Fourier transform yields amplitude and phase in the frequency domain. The former signifies style, while the latter encapsulates structure. Noise in raw images commonly resides in the phase component (Li et al. 2023), thus preliminary denoising aids the network in grasping color details more effectively. Fourier transform's decoupling facet enabling us separate processing of these distinct information.

Network Framework

The network architecture, depicted in Figure 3, processes the RAW image $\mathbf{R}^{H\times W\times 1}$. $\mathbf{R}$ undergoes packing and demosaicing to yield $\mathbf{R_{pack}}^{\frac{H}{2}\times\frac{W}{2}\times 4}$ and $\mathbf{R_{dem}}^{H\times W\times 3}$, respectively. PES, ARS, and CAS form the core components.

PES processes $\mathbf{R_{pack}}$ for spatial structure enhancement, yielding refined feature $\mathbf{F_P}^{H\times W\times C}$. The projection of $\mathbf{F_P}$ into RGB space involves convolutional steps, supervised with phase from the ground truth (GT) image. Simultaneously, $\mathbf{R_{dem}}$ is directed to ARS, optimizing amplitude to generate $\mathbf{F_A}^{H\times W\times C}$. This amplitude-focused information is then projected back into RGB space through convolutional projection, with supervision from GT image amplitudes.

Next, $\mathbf{F_P}$ is fed into CAS, leveraging $\mathbf{F_A}$'s amplitude to adjust color information via color adaptation blocks, as amplitude encodes image style. CAS generates the output image. Spatial loss and frequency domain loss govern final result supervision.

Key Components

Phase Enhance Subnet. The PES is specifically designed to enhance the spatial structure of the input images. Its primary

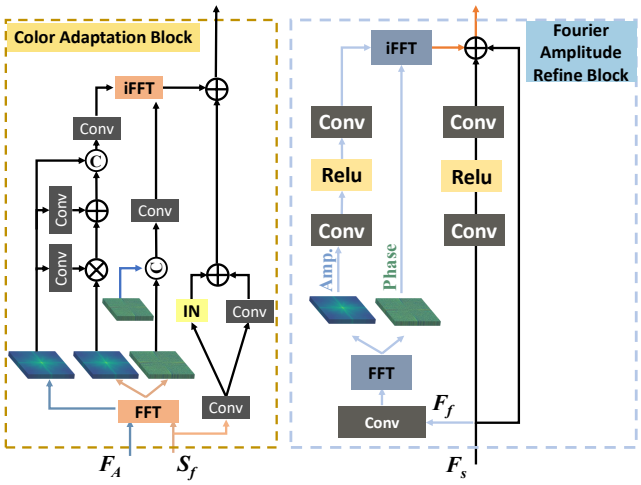

Figure 4: The left portion of the figure illustrates the Color Adaptation Block, while the right side showcases the Fourier Amplitude Refine Block.

function is to align the image phase with the ground truth, which facilitates fine-grained spatial structure learning. By optimizing the spatial structure, we can effectively reduce noise in the final output and generate finer image textures since image noise predominantly resides within the phase components. We start by packing the RAW image into four channels, a preprocessing step that enhances the image's suitability for spatial structure learning (Dai et al. 2020). The input to PES, denoted as $\mathbf{R_{pack}}$, undergoes processing, and we perform upsampling through PixelShuffle at the end. The process of PES can be described as follows:

$$\mathbf{F_p} = \phi(\mathbf{R_{pack}}), \tag{4}$$
$$\mathbf{Y_p} = Proj(\mathbf{F_p}) \tag{5}$$

Here, $Proj(.)$ represents the 1×1 convolution operator that projects $\mathbf{F_p}$ back to the RGB domain, and $\phi(.)$ corresponds to the PES.

The core component of PES is the Fourier Phase Refine Block (FPRB), depicted in Figure 4. It's worth noting that the main distinction between FPRB and the Fourier Amplitude Refine Block (FARB) is that FPRB deals with the phase, while FARB focuses on the amplitude. This module utilize the interaction between image phase information and spatial features, leading to efficient extraction of complex image details.

The operations within the FARB can be defined as follows, given the input features $\mathbf{F_f}$ and $\mathbf{F_s}$:

$$\mathbf{F_f^1} = Conv(\mathbf{F_f}), \tag{6}$$
$$\mathcal{A}(\mathbf{F_f^1}), \mathcal{P}(\mathbf{F_f^1}) = \mathcal{F}(\mathbf{F_f^1}), \tag{7}$$
$$\mathbf{F_f^2} = \mathcal{F}^{-1}(Conv(\mathcal{A}(\mathbf{F_f^1})), \mathcal{P}(\mathbf{F_f^1})), \tag{8}$$
$$\mathbf{F_{out}} = \mathbf{F_f^2} + Conv(\mathbf{F_s}) + \mathbf{F_s}. \tag{9}$$

Amplitude Refine Subnet. The architecture of the ARS closely resembles that of PES. ARS uses $\mathbf{R_{dem}}$ as input, ensuring that images of the same size are better suited for learning color mapping. In ARS, the primary objective is to optimize the image's amplitude to align it with the GT image. This amplitude component encodes the style of the image and encompasses global information, making color learning more robust to dataset misalignment. The utilization of global information is crucial for effective color learning, as images color are influenced by both local details and global features. Optimizing amplitude for color learning allows us to leverage the global attributes of the Fourier transform, enabling the capture of essential global information. . The FARB, as depicted in Figure 4, plays a central role in ARS for optimizing image amplitude. Its core structure is similar to the FPRB, with the key distinction being that FARB focuses on optimizing amplitude instead of phase.

The ARS process can be concisely expressed as follows:

$$\mathbf{F_A} = \gamma(\mathbf{R_{dem}}), \tag{10}$$
$$\mathbf{Y_A} = Proj(\mathbf{F_A}) \tag{11}$$

Here, $Proj(.)$ symbolizes a 1×1 convolution operator, projecting F_A back into the RGB domain and $\gamma(.)$ is ARS.

Color Adaptation SubNet. CAS serves as a pivotal component with the primary aim of fusing the refined spatial structures with accurate color information, thereby enabling precise adjustments to the color attributes within the feature maps. This sub-network adopts a multi-scale approach, injecting color features into the feature maps at different scales, resulting in targeted enhancement of both the overall color information and spatial structure across the entire image. Within CAS, a straightforward Unet architecture is employed, complemented by a custom-designed Color Adaptation Block (CAB), depicted in Figure 4, to modulate the feature map. We utilize amplitude features to modulate the style information within the feature map for color information is predominantly encoded by amplitude.

In CAB, the feature map is split into two branches. In the frequency branch, we apply Fourier transform to $\mathbf{F_A}$ and $\mathbf{S_f}$ independently. The amplitude of $\mathbf{F_A}$ is used to modulate the amplitude of $\mathbf{S_f}$ using the SFT (Wang et al. 2018) mechanism, and the phase components from both features are simply fused. After performing inverse Fourier transform, we obtain the frequency features. In the spatial domain branch, we employ the HINBlock (Chen et al. 2021) for efficient feature extraction, and subsequently, sum up the spatial and frequency features to complete the color adaptation process. The adapted feature map is then downsampled, serving as the $\mathbf{F_A}$ for the next scale.

Loss Function

Our comprehensive loss function comprises three essential components: Phase loss for PES, Amplitude loss for ARS, and the combined spatial and frequency loss for the final output.

PES and ARS are supervised using Phase Loss and Amplitude Loss. The Phase Loss $\mathcal{L}_{pha}$ is computed as the L1 norm between the phase $\mathcal{P}(\mathbf{Y_P})$ of the $\mathbf{Y_P}$ and the reference image GT's phase $\mathcal{P}(\mathbf{G})$. Similarly, the Amplitude Loss $\mathcal{L}_{amp}$ is calculated as the L1 norm between the amplitude $\mathcal{A}(\mathbf{Y_A})$ of $\mathbf{Y_A}$ and the reference image GT's amplitude

$\mathcal{A}(\mathbf{G})$.

$$\mathcal{L}_{pha} = ||\mathcal{P}(\mathbf{Y_P}) - \mathcal{P}(\mathbf{G})||_1, \quad (12)$$

$$\mathcal{L}_{amp} = ||\mathcal{A}(\mathbf{Y_A}) - \mathcal{A}(\mathbf{G})||_1. \quad (13)$$

The final network output, denoted as $\mathbf{Y}$, is constrained by L_{spa} (spatial loss) and L_{fre} (frequency domain loss). Regarding the spatial loss, we utilize a combined approach involving VGG (Johnson, Alahi, and Fei-Fei 2016), SSIM, and L1 losses. In the context of the frequency domain loss, we optimize the real and imaginary components of both $\mathbf{Y}$ and the reference image $\mathbf{G}$ after performing Fourier transform, which we represent as $R(Y)$, $R(G)$, $I(Y)$, and $I(G)$. The formulation of this loss component can be briefly expressed as follows:

$$\mathcal{L}_{spa} = \mathcal{L}_{vgg} + 0.5 * \mathcal{L}_{ssim} + \mathcal{L}_1, \quad (14)$$

$$\mathcal{L}_{fre} = ||R(Y) - R(G)||_1 + ||I(Y) - I(G)||_1 \quad (15)$$

Our comprehensive loss function is a weighted summation of the components outlined above:

$$\mathcal{L}_{total} = \mathcal{L}_{spa} + \alpha * \mathcal{L}_{fre} + \beta * \mathcal{L}_{pha} + \gamma * \mathcal{L}_{amp} \quad (16)$$

In our implementation, we set the weights α, β, and γ to 0.1 based on experience. Due to our incorporation of global information within the losses, our approach demonstrates reduced sensitivity to dataset misalignment issues.

Experiment

Datasets and Benchmark

We conducted evaluations on two distinct datasets: the ZRR dataset and the MAI dataset (Ignatov et al. 2021). The ZRR dataset involves mapping RAW images from the Huawei P20 camera to RGB images from a Canon camera. Meanwhile, the MAI dataset focuses on mapping Sony IMX586 Quad Bayer RAW images to Fuji camera RGB images. Notably, the RAW images in the ZRR dataset possess a bit width of 10 bits, whereas the MAI dataset RAW images have a bit width of 12 bits. The selection of the latter dataset allows us to assess the model's transferability. Given the substantial dissimilarity between the RAW images in these two datasets, we initially train the model on the ZRRdataset. To evaluate the model's transferability, we perform a one-epoch fine-tuning process on the MAI dataset. This approach serves to gauge how well the model can adapt to the differences in RAW image characteristics.

In our comparative analysis, we include state-of-the-art methods such as Pynet, AWNet, MWISP, MWISPGAN, LiteISP, and LWISP. We utilize reference evaluation metrics, including PSNR, SSIM, MS-SSIM (Wang, Simoncelli, and Bovik 2003), and LPIPS (Zhang et al. 2018). For more experiments results, please refer to the supplementary material

Implementation Details

We conducted our experiments utilizing the PyTorch framework on four Titan XP GPUs, encompassing a total of $3{*}10^4$ training iterations. Employing the Adam optimizer, we initially set the learning rate at $2 * 10^{-4}$, progressively halving it at every $1 * 10^4$ iterations to fine-tune the training process. Notably, for the ZRRdataset, we utilized a patch size of 448x488 for both training and testing. In contrast, for the MAI dataset, the patch size was set to 224x224 for the same purposes.

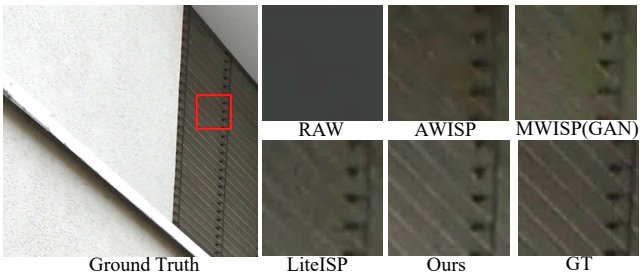

Figure 5: Experimental results from the ZRRdataset. our approach excels in capturing intricate texture details. RAW images are difficult to visualize due to dark local areas.

Comparison with State-of-the-Art Methods

Evaluation on Quantitative Metric. We conducted a comprehensive comparison of our proposed method against the SOTA approaches on both the ZRR and MAI datasets, as illustrated in the Table 1. For the MAI dataset, we employed a pre-trained model on the ZRR dataset and performed one epoch of fine-tuning for direct comparison. Additionaly,acknowledging the partial misalignment between the ground truth and input in the ZRR dataset, we employed the optical flow network (Sun et al. 2018) from LiteISP to align the test set and calculate the evaluation metrics. Due to unavailability of pretrained weights of Pynet and source code of LWISP, we based our evaluation of both methods solely on the metrics provided in their respective publication.

Our method achieved PSNR improvement of 0.08dB compared to the SOTA method on the ZRR dataset. Additionally, our method exhibited advancements in SSIM and LPIPS metrics. Remarkably, on the aligned dataset, our method achieved a noteworthy PSNR improvement of 0.17dB. This further reinforces the effectiveness of our style and structure decoupling approach and demonstrates the robustness of the global loss in handling data misalignment.

Moreover, our model exhibits strong transferability, surpassing the performance metrics of other methods on the MAI dataset. It is notable that the requirement for an additional optical flow estimation network during LiteISP training limits the transferability of the model, leading to bad result on the MAI dataset.

Evaluation on Qualitative Metric. In our qualitative experiments, we select three typical images from two datasets to comprehensively showcase the effectiveness of our method across three crucial aspects: image texture, color fidelity, and model transfer capability. As illustrated in Figure 5, we magnified local textures from the ground truth to emphasize the detail-capturing capabilities of various models. The visual comparison clearly indicates that our method, leveraging fine processing of spatial structures and the capacity

Methods	ZRR				ZRR(Align GT with RAW)				MAI			
	PSNR↑	SSIM↑	MS-SSIM↑	LPIPS↓	PSNR↑	SSIM↑	MS-SSIM↑	LPIPS↓	PSNR↑	SSIM↑	MS-SSIM↑	LPIPS↓
PyNet	21.19	0.7471	0.8620	0.1930	22.73	0.8451	/	0.1520	/	/	/	/
AWNET(raw)	21.42	0.7478	0.8609	0.1980	23.27	0.8542	0.9312	0.1510	23.95	0.8583	0.9508	0.1640
AWNET(demosaic)	21.53	0.7488	0.8614	0.2120	23.38	0.8497	0.9297	0.1640	24.03	0.8670	0.9525	0.1340
MWISP	21.42	0.7544	0.8654	0.2130	23.07	0.8479	0.9255	0.1650	24.24	0.8558	0.9491	0.1280
MWISP-GAN	21.16	0.7317	0.8578	**0.1580**	22.80	0.8285	0.9234	0.1340	24.34	0.8568	0.9234	0.1150
LiteISP	21.55	0.7487	0.8596	0.1870	23.76	0.8730	0.9450	0.1330	23.11	0.7941	0.9250	0.1990
LWISP	21.57	/	0.8622	/	/	/	/	/	/	/	/	/
Ours	**21.65**	**0.7546**	**0.8660**	0.1820	**23.93**	**0.8744**	**0.9461**	**0.1240**	**24.99**	**0.8820**	**0.9594**	**0.0850**

Table 1: Quantitative comparison on three datasets. Best results are highlighted by **bold**. ↑ indicates that the larger the value, the better the performance, and ↓ indicates that the smaller the value, the better the performance.

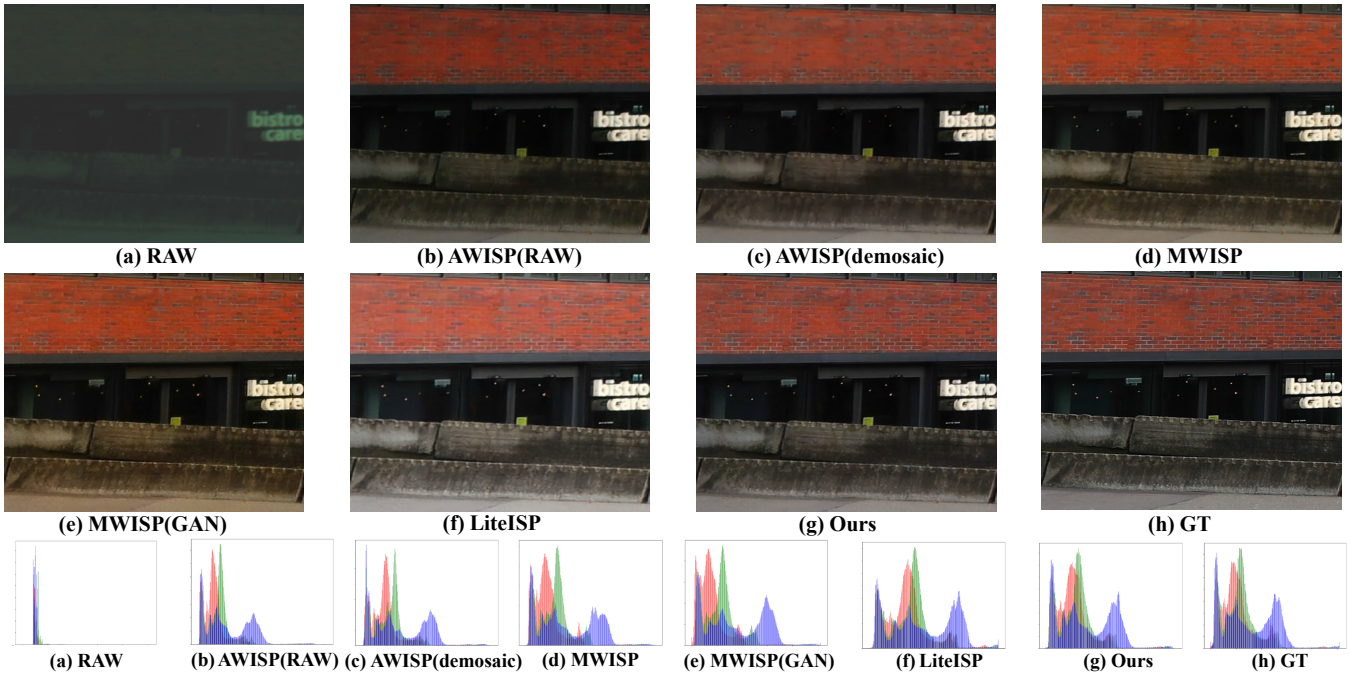

Figure 6: The results image from ZRRdataset. The last row showcase the color histogram of the image. Our method has the closest color to GT image.

Figure 7: The Amplitude and Phase feature map of our methods.

to capture high-frequency information in the frequency domain, consistently generates more refined results, in contrast to other methods that struggle to capture such textures, even leading to artifacts.

In Figure 6, we present a comprehensive comparison of color fidelity, along with accompanying color histograms for each image. Analyzing the histograms reveals that our method's results closely align with the ground truth, signifying our superior color reproduction. Finally, we demonstrate the model's transferability by comparing images from the MAI dataset in Figure 8. The comparison highlights that alternative methods exhibit incorrect color mappings or spatial structures, while our method consistently aligns more closely with the ground truth in terms of both color and spa-

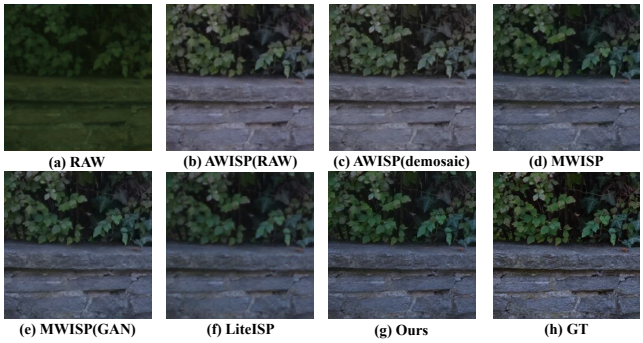

Figure 8: MAI dataset result image. Our method showcases the superiortransferability, closely aligned with the Ground Truth.

Configuration	ZRR			
	PSNR	SSIM	MS-SSIM	LPIPS
FourierISP w/o Phase	21.47	0.7413	0.8581	0.199
FourierISP w/o Amplitude	21.37	0.7488	0.8592	0.187
FourierISP w/ ColorHist	21.13	0.7476	0.8581	0.193
FourierISP	**21.65**	**0.7546**	**0.8660**	**0.182**

Table 2: The results of the ablation experiments conducted on the ZRRdataset

Config	Params(M)	ZRR			
		PSNR	SSIM	MS-SSIM	LPIPS
16 channels	2.75	21.64	0.7534	0.8609	0.186
48 channels	24.6	**21.67**	**0.7548**	**0.8669**	**0.180**
24 channels	6.17	21.65	0.7546	0.8660	0.182

Table 3: The results of the ablation experiments conducted on the ZRRdataset

tial information.

Ablation Experiments

We conducted multiple ablation experiments on the ZRR dataset to validate our method, and conducted experiments on multiple dimensions such as model structure, loss function, and model parameter quantity.

Model Structure. Our model is thoughtfully designed to capture both spatial structures and style features, with dedicated modules in PES and ARS, respectively. In this ablation experiments, we removed certain components from the network's architecture. Specifically, in the PES configuration, we omitted the frequency branch of FPRB and the corresponding phase loss, and similarly, in ARS, we excluded the frequency branch of FARB, the CAB, and the Amplitude Loss. We directly concatenate $\mathbf{F_A}$ with $\mathbf{F_P}$ into CAS. The results of ablation experiments are presented in the first and second rows of the Table 2. Removal of any module leads to a significant drop in both evaluation metrics. Specifically, when Phase information is omitted, SSIM indicators decrease notably, indicating severe damage to the result image spatial structure. Conversely, the removal of Amplitude information primarily affects the PSNR indicator.

Loss Function. In the examination of the loss function, we conducted empirical investigations by substituting the AmplitudeLoss in AmplitudeNet with ColorHistLoss to assess its influence on the final results. Prior research (Afifi, Brubaker, and Brown 2021) has utilized ColorHistLoss to approximate the color distribution of an image towards a reference image. Our findings shown in the third row of Talbe 2 indicate that when replacing AmplitudeLoss with ColorHistLoss, the model's performance indicators exhibited a decline, emphasizing that the coarse-grained histogram partitioning approach is not well-suited for the Raw to RGB task.

Model Parameters. In addition, we conducted a comprehensive investigation into the influence of parameter quantities on model performance. Our initial experiment employed a base channel count of 24, while in the ablation study, we explored channel counts of 16 and 48 to assess their impact on model effectiveness. As depicted in the results table 3, a direct correlation is observed between the number of parameters and the model's performance, with larger parameter counts leading to improved outcomes. While constraining the channel count to 16 effectively reduces the model's parameter count, and the achieved PSNR index remains comparable to our existing methods, the visual results fall short of the quality attained with our current settings. our model possesses half the number of parameters compared to LiteISP and only one-sixth of MWISP, placing it on par with LWISP. This highlights our model's equilibrium between the quantity of model parameters and evaluation metrics.

Visualization of Feature Maps

To demonstrate the distinctive capabilities of our subnetworks, we conducted feature map visualizations shown in Figure 7. Specifically, we projected $\mathbf{F_A}$ and $\mathbf{F_P}$ into the RGB space, yielding $\mathbf{Y_A}$ and $\mathbf{Y_P}$, respectively. The visual analysis reveals that PES adeptly captured the spatial structure, while ARS faithfully preserved the image's style attributes.

Remarkably, $\mathbf{Y_A}$ captures the overall style of the image. However, it tends to lack granularity in representing local details. On the other hand, $\mathbf{Y_P}$ focuses on preserving the structural details and finer local information of the image but lacks color information.

This visualization underscores the effectiveness of our method in handling both structural and style information.

Conclusion

This paper introduces a novel approach for RAW to RGB mapping, leveraging the power of Fourier transform to disentangle image style and structure. By processing these two aspects independently, our method achieves remarkable accuracy in color reproduction and texture preservation. Our proposed Fourier ISP framework features Phase Enhance Networks, Amplitude Refine Networks, and Color Adaptation Networks, enabling separate learning of style and structure, followed by a coherent integration for the final output. Extensive quantitative and qualitative experiments on multiple datasets demonstrate the superiority of our approach over state-of-the-art methods.

Acknowledgements

This work was supported by the Natural Science Foundation of Anhui Province (No.2208085MC57), and HFIPS Director's Fund, Grant No.2023YZGH04.

References

Afifi, M.; Brubaker, M. A.; and Brown, M. S. 2021. Histogan: Controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7941–7950.

Chen, H.; and Ma, K. 2022. LW-ISP: A Lightweight Model with ISP and Deep Learning. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, 148. BMVA Press.

Chen, L.; Lu, X.; Zhang, J.; Chu, X.; and Chen, C. 2021. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 182–192.

Cheng, S.; Wang, Y.; Huang, H.; Liu, D.; Fan, H.; and Liu, S. 2021. Nbnet: Noise basis learning for image denoising with subspace projection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4896–4906.

Chi, L.; Jiang, B.; and Mu, Y. 2020. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33: 4479–4488.

Dai, L.; Liu, X.; Li, C.; and Chen, J. 2020. Awnet: Attentive wavelet network for image isp. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 185–201. Springer.

Hu, L.; Chen, H.; and Allebach, J. P. 2022. Joint multi-scale tone mapping and denoising for HDR image enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 729–738.

Ignatov, A.; Chiang, C.-M.; Kuo, H.-K.; Sycheva, A.; and Timofte, R. 2021. Learned smartphone isp on mobile npus with deep learning, mobile ai 2021 challenge: Report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2503–2514.

Ignatov, A.; Timofte, R.; Zhang, Z.; Liu, M.; Wang, H.; Zuo, W.; Zhang, J.; Zhang, R.; Peng, Z.; Ren, S.; et al. 2020. Aim 2020 challenge on learned image signal processing pipeline. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 152–170. Springer.

Ignatov, A.; Van Gool, L.; and Timofte, R. 2020. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 536–537.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 694–711. Springer.

Li, C.; Guo, C.; Zhou, M.; Liang, Z.; Zhou, S.; Feng, R.; and Loy, C. C. 2023. Embedding Fourier for Ultra-High-Definition Low-Light Image Enhancement. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Liang, Z.; Cai, J.; Cao, Z.; and Zhang, L. 2021. Cameranet: A two-stage framework for effective camera isp learning. *IEEE Transactions on Image Processing*, 30: 2248–2262.

Liu, L.; Jia, X.; Liu, J.; and Tian, Q. 2020. Joint demosaicing and denoising with self guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2240–2249.

Liu, P.; Zhang, H.; Zhang, K.; Lin, L.; and Zuo, W. 2018. Multi-level wavelet-CNN for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 773–782.

Ramanath, R.; Snyder, W.; Yoo, Y.; and Drew, M. 2005. Color image processing pipeline. *IEEE Signal Processing Magazine*, 22(1): 34–43.

Schwartz, E.; Giryes, R.; and Bronstein, A. M. 2018. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE Transactions on Image Processing*, 28(2): 912–923.

Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8934–8943.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159.

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 606–615.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, 1398–1402. Ieee.

Yang, K.-F.; Cheng, C.; Zhao, S.-X.; Yan, H.-M.; Zhang, X.-S.; and Li, Y.-J. 2023. Learning to adapt to light. *International Journal of Computer Vision*, 131(4): 1022–1041.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Z.; Wang, H.; Liu, M.; Wang, R.; Zhang, J.; and Zuo, W. 2021. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4348–4358.

Zhou, J.; and Glotzbach, J. 2007. Image pipeline tuning for digital cameras. In *2007 IEEE International Symposium on Consumer Electronics*, 1–4. IEEE.

Zhou, M.; Huang, J.; Guo, C.-L.; and Li, C. 2023. Fourmer: An Efficient Global Modeling Paradigm for Image Restoration. In *International Conference on Machine Learning*, 42589–42601. PMLR.

Zhou, M.; Huang, J.; Yan, K.; Yu, H.; Fu, X.; Liu, A.; Wei, X.; and Zhao, F. 2022. Spatial-frequency domain information integration for pan-sharpening. In *European Conference on Computer Vision*, 274–291. Springer.