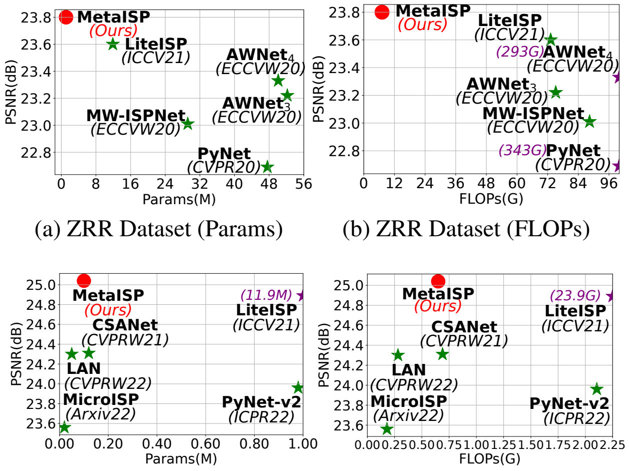
MetaISP: Efficient RAW-to-sRGB Mappings with Merely 1M Parameters

Zigeng Chen[1], **Chaowei Liu**[1], **Yuan Yuan**[2], **Michael Bi Mi**[2], **Xinchao Wang**[1]*
[1]National University of Singapore
[2]Huawei Technologies Ltd
zigeng99@u.nus.edu, e1011116@u.nus.edu, xinchao@nus.edu.sg

Abstract

State-of-the-art deep ISP models alleviate the dilemma of limited generalization capabilities across heterogeneous inputs by increasing the size and complexity of the network, which inevitably leads to considerable growth in parameter counts and FLOPs. To address this challenge, this paper presents MetaISP - a streamlined model that achieves superior reconstruction quality by adaptively modulating its parameters and architecture in response to diverse inputs. Our rationale revolves around obtaining corresponding spatial and channel-wise correction matrices for various inputs within distinct feature spaces, which assists in assigning optimal attention. This is achieved by predicting dynamic weights for each input image and combining these weights with multiple learnable basis matrices to construct the correction matrices. The proposed MetaISP makes it possible to obtain the best performance while being computationally efficient. SOTA results are achieved on two large-scale datasets, e.g. 23.80dB PSNR on ZRR, exceeding the previous SOTA **0.19dB** with only **9.2%** of its parameter count and **10.6%** of its FLOPs; 25.06dB PSNR on MAI21, exceeding the previous SOTA **0.17dB** with only **0.9%** of its parameter count and **2.7%** of its FLOPs.



(a) ZRR Dataset (Params) (b) ZRR Dataset (FLOPs)

(c) MAI21 Dataset (Params) (d) MAI21 Dataset (FLOPs)

Figure 1: Our MetaISP achieves the best performance on two large-scale datasets and meanwhile being impressive computational efficiency and significantly lightweight.

1 Introduction

Over the past decade, DSLR cameras have increasingly been supplanted in various application scenarios due to their lack of portability. This shift has intensified the interest in achieving high-quality sRGB images without relying on large sensors and lenses. As sensor size constraints reach their maximum, researchers have turned their attention to improving the image signal processing (ISP) pipeline, which aims to reconstruct high-quality sRGB images from raw sensor images [Ignatov *et al.*, 2022a]. The conventional ISP pipeline consists of multiple manually designed modules. Unfortunately, the accumulation of errors from each processing module gradually degrades the overall reconstruction quality of the sRGB image [Liang *et al.*, 2021b].

To address this issue of accumulated reconstruction errors, the concept of an end-to-end deep ISP model has been introduced in the literature [Schwartz *et al.*, 2018], which integrates the entire ISP pipeline into one single unified learnable model. Despite the encouraging performance in RAW-to-sRGB tasks, these existing end-to-end deep ISP models present two distinctive challenges:

- **(1) Diverse Conditions.** The diversity in environmental conditions and camera settings under which images are captured makes it challenging for the model to establish accurate mappings for every image pair;
- **(2) High Efficiency Requirement.** Since the deep ISP model is commonly deployed on resource-constrained devices, prioritizing efficiency and lightweight design becomes crucial to ensure its practical value.

To attain remarkable reconstruction quality under diverse conditions, certain heavy ISP models [Ignatov *et al.*, 2020a; Ignatov *et al.*, 2020b; Kim *et al.*, 2020; Dai *et al.*, 2020; Liang *et al.*, 2021b; Zhang *et al.*, 2021] have been proposed. Yet, these models all come with complex struc-

*Corresponding author.

tures and a substantial parameter count. To be suitable for resource-constrained devices, some lightweight ISP models [Raimundo *et al.*, 2022; Ignatov *et al.*, 2023b; Hsyu *et al.*, 2021; Ignatov *et al.*, 2023a; Ignatov *et al.*, 2022a] have also been introduced. Unfortunately, these lightweight models all inevitably compromise the quality of the reconstructed sRGB image. As such, it remains a challenging problem in the literature to generate sRGB images of comparable or superior quality to those produced by heavy models, while utilizing significantly fewer parameters and FLOPs like those lightweight models.

In this paper, we introduce MetaISP, an adaptive network for RAW to sRGB mappings that offers a lightweight architecture and efficient inference while consistently producing high-quality outputs. Our rationale comes from the substantial variations of the mapping relationship between RAW and sRGB images under diverse conditions. Motivated by this observation, we propose an elaborated sample-wise dynamic network that adaptively adjusts its parameters and architecture in response to diverse inputs. Unlike existing models with fixed parameters, our method allows for automatic and flexible learning of accurate RAW to sRGB mappings under diverse conditions. Consequently, it presents significant facilitators to the model training process and enhances the model's ability to generalize. As depicted in Figure 1, our approach achieves high-quality sRGB images while significantly reducing the number of parameters and required FLOPs.

The key element of our MetaISP is a U-Net-like architecture comprising two novel blocks: the meta channel correction block (MCCB) and the meta spatial correction block (MSCB). MCCB obtains stronger parameter prediction ability and improves the learning capability for the relationship between feature map channels by predicting the content-dependent weights of the basis matrices. While MSCB innovatively calculates the cosine similarity between each pixel in the feature map and the adaptive correction vector to get the spatial attention map efficiently and achieve stronger representation ability. Additionally, the meta shallow feature extraction block (MFEB) and the meta reconstruction block (MRB) are also proposed to further improve the quality of the sRGB image. MFEB realizes flexible sample-wise feature extraction by predicting dynamic weights for five predefined convolution kernels based on the input RAW image. Additionally, MRB enhances global alterations by utilizing the output from U-Net's bottom layer to guide the final image generation process.

Our contribution is a novel deep ISP model that achieves high-quality sRGB images while significantly reducing computational complexity. This is achieved by four complementary and lightweight dynamic blocks entitled MCCB, MSCB, MFEB, and MRB, which adaptively adjust model parameters and structure based on input images. As shown in Figure 1, extensive experimental results on two large-scale datasets demonstrate the superiority of our MetaISP: 23.80dB PSNR on ZRR [Ignatov *et al.*, 2020b], exceeding the previous SOTA 0.19dB with only 9.2% of its parameter count and 10.6% of its FLOPs; 25.06dB PSNR on MAI21 [Ignatov *et al.*, 2022b], exceeding the previous SOTA 0.17dB with only 0.9% of its parameter count and 2.7% of its FLOPs.

2 Related Work

Recent approaches in Image Signal Processing (ISP) aim to enhance image quality by leveraging deep learning to transform Bayer RAW format images into high-quality RGB images. These methods formulate ISP tasks as end-to-end processes, replacing the traditional sequential modules with a single model.

Schwartz *et al.*[Schwartz *et al.*, 2018] pioneered this approach by introducing DeepISP, which utilizes Convolutional Neural Networks (CNNs) to create a fully integrated ISP pipeline. Building upon this, Ignatov *et al.* [Ignatov *et al.*, 2020b] developed Pynet, a network designed around the pyramid architecture, thereby further advancing the receptive field. This development encouraged many researchers to adapt Unet-based models for this task. Zhang *et al.* [Ignatov *et al.*, 2020a] proposed MW-ISPNet, a model that first uses a wavelet transform to replace the downsampling and upsampling stages in the deep ISP model. Additionally, Dai *et al.* introduced AWNet [Dai *et al.*, 2020], incorporating wavelet transform and global context attention to enhance image quality by providing the model with a larger receptive field. Raimundo *et al.* [Raimundo *et al.*, 2022] proposed LAN, a model integrating spatial attention to capture spatial information within images. Liang *et al.*[Liang *et al.*, 2021b] suggested CamerNet, arguing that ISP should be split into two relatively uncorrelated segments. They deployed two separate U-Net models to handle restoration and enhancement, respectively. LiteISP [Zhang *et al.*, 2021] achieves better results with a lighter model by solving the problem of misalignment between RAW and sRGB images in the dataset.

In an effort to boost ISP performance on mobile platforms, [Ignatov *et al.*, 2022a] proposed PynetV2, a lightweight and efficient model that expands upon the concept of channel attention. [Hsyu *et al.*, 2021] proposed CSANet, which employs a dual attention module that leverages both channel and spatial attention. [Ignatov *et al.*, 2023a] further innovated with the proposal of MicroISP, which operates at the original scale, thereby reducing the GPU memory requirements. This approach enables the processing of larger images and ensures faster operational speed.

These state-of-art ISP algorithms still suffer from one major flaw: in order to accommodate the diversity of input RAW images derived from real-world environments, these methods often resort to expanding the model's capacity to enhance its generalizability. However, this invariably leads to an escalation in computational cost. In contrast to existing methods, our approach eliminates the dependency on high-cost models, while still achieving high-quality generation of RGB images.

3 Proposed Method

Our primary goal is to develop a deep ISP model for high-quality sRGB image restoration from camera sensor outputs, with reduced parameters and FLOPs. To achieve exceptional performance with low computational complexity, we propose MetaISP, which utilizes a dynamic network concept. In this section, we introduce MetaISP's overall architecture and four key components: (a) meta channel correction block (MCCB), (b) meta spatial correction block (MSCB), (c) meta shallow

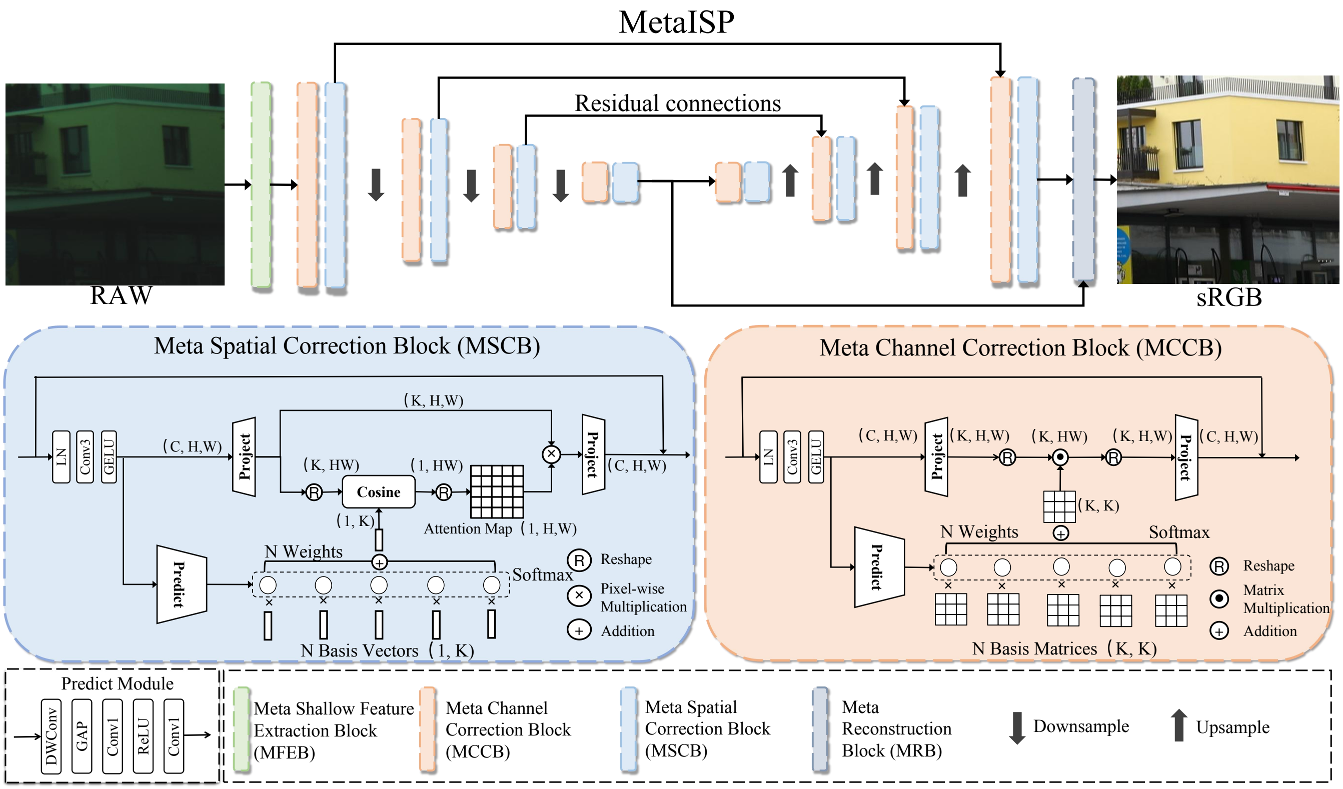


Figure 2: Overall architecture of MetaISP, the architecture of meta channel correction block (MCCB) and meta spatial correction block (MSCB).

feature extraction block (MFEB), and (d) meta reconstruction block (MRB).

3.1 Overview

As illustrated in Figure 2, the overall architecture of MetaISP follows a U-shaped encode-decoder structure. The U-net architecture enables the model to capture both low-level and high-level image features, enabling robust representation learning. Given a 4-channel RAW image $I_{raw} \in \mathbb{R}^{4\times H/2\times W/2}$, MetaISP first employ MFEB to adaptively extract shallow feature $F_0 \in \mathbb{R}^{C\times H/2\times W/2}$, where $H/2 \times W/2$ denotes the spatial resolution and C is the number of channels. Then the shallow features F_0 are fed into a 4-level U-net for multi-scale deep feature extraction and reconstruction. Each level of the U-net contains only two correction blocks which are composed of proposed MCCB and MSCB. From top to bottom levels, the number of feature map channels are $\{C, C, C, 2C\}$. The model benefits from a comparatively smaller number of channels and blocks, resulting in a significant reduction in computational complexity. The downsampling and upsampling of the feature maps are achieved using discrete wavelet transformation [Liu *et al.*, 2018] and a 3×3 convolution layer. The residual connections are implemented through per-pixel addition. Finally, the feature map of the bottom level $F_2 \in \mathbb{R}^{2C\times H/16\times W/16}$ and the output of the U-net $F_1 \in \mathbb{R}^{C\times H/2\times W/2}$ are fed into MRB to reconstruct the sRGB image $I_{rgb} \in \mathbb{R}^{3\times H\times W}$.

3.2 Meta Channel Correction Block (MCCB)

Channel attention [Hu *et al.*, 2018] is a simple dynamic parameter block that is commonly used in many low-level vision tasks including deep ISP to capture channel-wise dependencies and highlight informative features while suppressing irrelevant or redundant ones. By incorporating the channel attention mechanism into CNN architectures, models can adaptively assign different importance levels to channels and dynamically adjust their contributions during feature extraction. It realizes computational efficiency and brings global information to the feature map. However, its linear transformation through simple channel-wise multiplication with attention weights $M \in \mathbb{R}^{C\times 1}$ exhibits limited fitting capacity, while its simplistic attention prediction layers may lead to suboptimal attention weights. Consequently, it becomes necessary to continually increase the depth and width of the channel attention block to compensate for these drawbacks. To further improve the ability to learn inter-channel relationships, Restormer [Zamir *et al.*, 2022] proposed a more powerful and dynamic method based on the self-attention (SA) mechanism [Vaswani *et al.*, 2017]. It calculates the response of a specific channel by performing a weighted sum across all other channels, thereby generating an attention map $M \in \mathbb{R}^{C\times C}$. Nevertheless, generating the attention map involves computing cross-covariance across channels, which is computationally more expensive compared to CNN-based channel attention.

To achieve the high performance of self-attention while maintaining low computational complexity similar to CNN, we propose MCCB. The architecture of MCCB is shown in Figure 2. The block begins with layer normalization [Ba *et al.*, 2016], which enhances the stability of the training process and enables the utilization of larger learning rates. The following 3×3 convolution and GELU [Hendrycks and Gimpel, 2016] aggregate the local spatial context. Then the feature map $X \in \mathbb{R}^{C\times H\times W}$ is projected to a high-dimensional space $Y \in \mathbb{R}^{K\times H\times W}$ to obtain richer features. Unlike MDTA and channel attention, which directly calculate the weights of each channel, MCCB joint learns several basis matrices $\{\phi_n\}_{n=1,\ldots,N}$ together with a simple prediction branch p which predicts weights $\{w_n\}_{n=1,\ldots,N} = p(X)$ for each basis matrices. For an input feature map $X \in \mathbb{R}^{C\times H\times W}$, the adaptive correction matrix $M \in \mathbb{R}^{K\times K}$ is obtained as:

$$M = \sum_{n=1}^{N} w_n(x)\phi_n. \tag{1}$$

Next, we reshape the high-dimensional feature map and perform matrix multiplication with the input-adaptive correction matrix, similar to the self-attention mechanism. The transformation process is defined as:

$$\widehat{R} = R * M, \tag{2}$$

where $\widehat{R} \in \mathbb{R}^{HW\times K}$ and $R \in \mathbb{R}^{HW\times K}$ are the input and output reshaped feature maps, $M \in \mathbb{R}^{K\times K}$ is the input-adaptive correction matrix, $*$ stands for matrix multiplication. Finally, we project the transformed feature map $\widehat{Y} \in \mathbb{R}^{K\times H\times W}$ back to its original dimension $\widehat{X} \in \mathbb{R}^{C\times H\times W}$ and establish a residual connection [He *et al.*, 2016] with the input $X \in \mathbb{R}^{C\times H\times W}$.

Our MCCB utilizes only a few basis matrices to transform the feature maps and employs dynamic soft weights to achieve content-adaptive channel transformation. This mechanism enables us to leverage the same low computational complexity as channel attention while effectively learning complex inter-channel relationships, similar to the transformer model.

3.3 Meta Spatial Correction Block (MSCB)

In addition to channel attention, spatial attention is another dynamic parameter mechanism that focuses on capturing and emphasizing relevant spatial regions in an image. One common approach in spatial attention is to utilize convolutional neural networks (CNNs) to learn attention maps that highlight relevant spatial regions. These attention maps are then used to modulate the feature representations in subsequent layers, enabling the network to focus on discriminative regions while suppressing irrelevant or noisy information. The CNN-based spatial attention leverages parameter sharing across spatial locations, enabling efficient processing and reducing computational complexity. However, it has a limited receptive field, thus preventing it from modeling long-range pixel interactions. While the self-attention-based spatial attention [Chen *et al.*, 2021; Liang *et al.*, 2021a; Wang *et al.*, 2022] has exceptional capabilities in global context modeling but suffers from high computational complexity. The complexity of the attention block is quadratic in relation to the input size, making it impractical to apply to high-resolution images in ISP tasks. To simultaneously achieve the low computational complexity of CNN and the global context modeling capabilities of self-attention in spatial attention, we propose the MSCB.

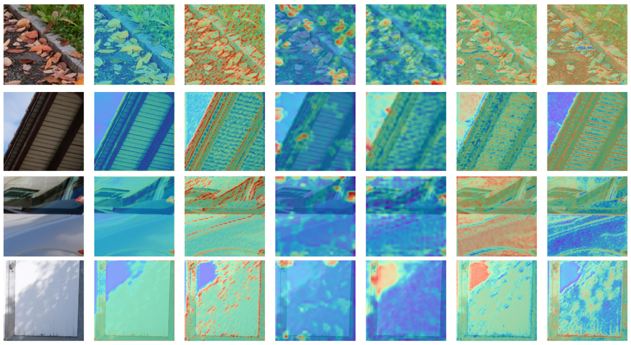


Figure 3: The visualizations of attention maps in MSCB, ordered from left to right across U-Net layers.

As is shown in Figure 2, MSCB shares a comparable overall architecture with MCCB. We simultaneously learn multiple basis vectors $\{\varepsilon_n\}_{n=1,\ldots,N}$ along with a straightforward prediction branch, denoted as p, which predicts dynamic weights $\{w_n\}_{n=1,\ldots,N} = p(X)$ for each basis vectors. The adaptive correction vector $V \in \mathbb{R}^{1\times K}$ is obtained through weighted addition of multiple basis vectors:

$$V = \sum_{n=1}^{N} w_n(x)\varepsilon_n. \tag{3}$$

We assign attention weights to each pixel in the feature map by computing the cosine similarity between each pixel and the correction vector. The transformation process is defined as:

$$\widehat{Y} = Y \otimes S, \tag{4}$$

$$S_{w,h} = \frac{Y_{w,h} * V}{|Y_{w,h}| \times |V|}, \tag{5}$$

where $\widehat{Y} \in \mathbb{R}^{K\times H\times W}$ and $Y \in \mathbb{R}^{K\times H\times W}$ are the input and output high-dimensional feature maps, $V \in \mathbb{R}^{1\times K}$ is the sample-wise adaptive correction vector, $S \in \mathbb{R}^{1\times H\times W}$ is the spatial attention map, $\otimes$ stands for pixel-wise multiplication and $*$ stands for dot product. To normalize the attention weights, we compute the cosine similarity, ensuring that all values fall within the range of 0 to 1. This normalization process helps to achieve smoother training and reduce the excessive influence of attention weights on a few pixels. Pixels that exhibit higher similarity to the correction vector will be assigned higher attention weights, while pixels that show less similarity to correction vector will be assigned lower attention weights. MSCB achieves linear complexity $O\,(HW)$ while effectively utilizing the global information of the feature map.

The visualization of attention maps in MSCB is depicted in Figure 3, where the model leverages the attention mechanism to discern intrinsic spatial correlations. Attention maps at different levels reveal varying inherent relationships.

3.4 Meta Feature Extraction Block (MFEB)

Considering the varied demands for high-frequency (detail-focused) and low-frequency (color and illumination) information restoration across images, we advocate for a dynamic,

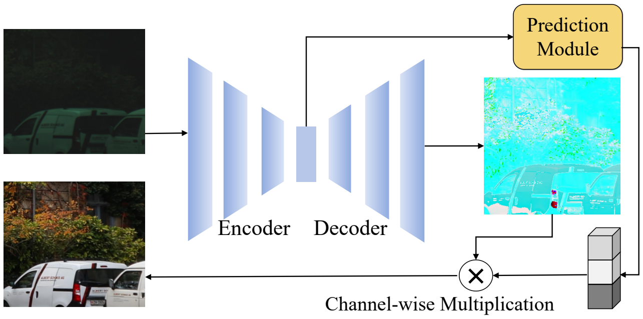


Figure 4: Architecture of the proposed MRB.

image-specific feature extraction. MFEB is chiefly partitioned into three components. We first implement a 3×3 standard convolution for baseline performance. Next, to gather global information, we apply pre-defined kernel functions such as the box filter and Gaussian filter. Lastly, we use Sobel-X, Sobel-Y, and Laplacian kernel functions for local feature extraction. A prediction network is also incorporated to generate weights for each branch based on the input image. The predicted weights $\{W_n\}_{n=1,\ldots,6}$ indicate the significance of each feature extraction branch, and $\{F_n\}_{n=1,\ldots,6}$ denote features extracted from each branch. The formula below depicts the weighted sum for each instance.

$$X_{out} = \sum_{n=1}^{6} W_n \times F_n. \tag{6}$$

3.5 Meta Reconstruction Block (MRB)

The crux of Learnable ISP tasks lies in color mapping. Modern image restoration typically employs U-Net-shaped networks to extract and decode multi-scale information for image restoration. Still, aspects like lighting and color require sequential decoding, typically at the U-Net's lowest layer. To address this, we introduce the meta reconstruction block. As shown in Figure 4, MRB uses the output from the U-Net's base layer to guide final image reconstruction.

$$sRGB = F_{reconstruct}\left(X_h\right) \otimes F_{channel}\left(X_l\right). \tag{7}$$

In the preceding equation, X_h and X_l depict the outputs from the top and bottom layers of the U-Net, respectively. F_r signifies a series of 3×3 convolutions with pixel shuffling, and $F_{channel}$ comprises standard 1×1 convolutions and a global average pooling, yielding a $C \times 1 \times 1$ output that captures inter-channel correlations. Finally, $\otimes$ indicates channel-wise multiplication.

4 Experiments

4.1 Implementation Details

Datasets. We conduct experiments on two publicly available RAW to sRGB datasets: the Zurich RAW to RGB (ZRR) dataset [Ignatov *et al.*, 2020b] and the MAI2021 dataset [Ignatov *et al.*, 2022b].

ZRR dataset is a large-scale dataset consisting of 48k RAW-sRGB pairs of size 448 × 448. We follow the official division that 46.8k are used for training and 1.2k are used for testing. To further address the misalignment issue in the ZRR dataset, we utilized a global color mapping (GCM) module [Zhang *et al.*, 2021] in combination with a pre-trained PWC-Net [Sun *et al.*, 2018] to warp the sRGB images with the corresponding RAW images. In contrast to the joint training strategy of [Zhang *et al.*, 2021], we pre-train a GCM module and utilize its output to align the sRGB images before training the ISP model for fairness and certainty in our comparisons.

We also used the MAI21 dataset which consists of 24k RAW-sRGB pairs of size 256 × 256. The dataset was then randomly divided into two parts, with 23k samples allocated for training and 1k samples reserved for testing.

Training Details. Our model is implemented in PyTorch [Paszke *et al.*, 2019] and trained on 4 Nvidia Titan X GPUs with a batch size of 32. The parameters of the network are optimized with ADAM [Kingma and Ba, 2014] algorithm.

For the ZRR dataset, all the training images were augmented by random horizontal and vertical flipping during the training. Our MetaISP consists of a four-level U-Net architecture, with channel numbers of 32, 32, 32, and 64 from top to bottom. The model undergoes a two-stage training process. First, the model is trained for 80 epochs with an initial learning rate of $5e^{-4}$ which is decayed to half after 50 epochs. The loss function is a combination of VGG-based perceptual loss [Johnson *et al.*, 2016], SSIM loss [Wang *et al.*, 2004] and Charbonier loss [Zhang *et al.*, 2018]:

$$\mathcal{L}_{Stage1} = 0.25 \cdot \mathcal{L}_{Char} + \mathcal{L}_{SSIM} + \mathcal{L}_{VGG}. \tag{8}$$

Next, the model is fine-tuned for an additional 5 epochs with a learning rate of $2e^{-5}$. Only MSE loss and SSIM loss are employed for final tone adjustments and edge rendering:

$$\mathcal{L}_{Stage2} = 0.5 \cdot \mathcal{L}_{MSE} + \mathcal{L}_{SSIM}. \tag{9}$$

For the MAI21 dataset, no data augmentation methods were employed throughout the training phase. We further reduce the depth and width of MetaISP to make it more lightweight. The training process is similar to the ZRR dataset.

4.2 Experimental Results

ZRR Dataset. Our proposed model, MetaISP, was benchmarked against four cutting-edge models on the ZRR dataset, demonstrating superior performance in all metrics, including PSNR, SSIM and ΔE (see Table 1). Notably, the MetaISP's excellence was achieved with significantly fewer parameters and computational resources, requiring only approximately 8.6% of the parameters and 9.5% of the Flops utilized by LiteISP [Zhang *et al.*, 2021]. Further corroborating its superiority, MetaISP synthesizes output images of an exceptional quality that are rich in detail and exhibit superb restoration of global information, such as illumination and color. This qualitative superiority is clearly illustrated in Figure 5.

In Table 2, we provide a detailed comparison of the actual memory requirements, running latency, and frames per second (FPS) across various image resolutions to offer a comprehensive evaluation of efficiency. Our model demonstrates significant advantages in efficiency over the SOTA

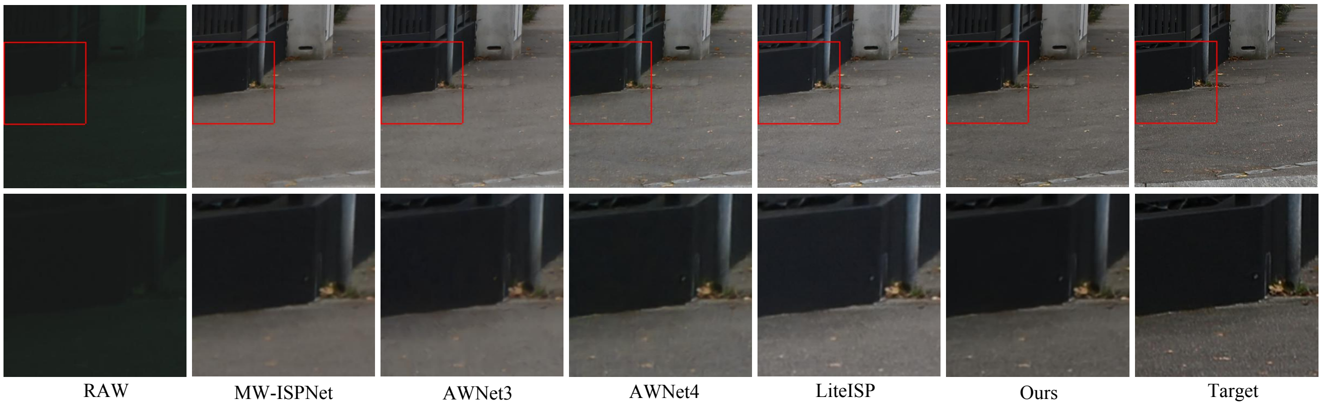


Figure 5: Visual comparison of our method with state-of-the-art approaches on the ZRR dataset [Ignatov *et al.*, 2020b]. Our MetaISP demonstrates superior performance in both global color fidelity and local detail preservation.

Dataset	Methods	Params(M) ↓	FLOPs(G) ↓	PSNR(dB) ↑	SSIM ↑	ΔE ↓	Extra Data
ZRR	PyNet[Ignatov *et al.*, 2020b]	47.6	343	22.69	0.8490	6.498	×
	AWNet$_3$ [Dai *et al.*, 2020]	52.2	75	23.22	0.8516	6.247	×
	AWNet$_4$ [Dai *et al.*, 2020]	50.1	293	23.33	0.8473	6.397	×
	MW-ISPNet [Ignatov *et al.*, 2020a]	29.2	88	23.01	0.8453	6.216	×
	LiteISP [Zhang *et al.*, 2021]	11.9	73	23.61	0.8672	6.079	×
	Ours (MetaISP)	**1.1**	**8**	**23.80**	**0.8758**	**5.881**	×
MAI21	CSANet [Hsyu *et al.*, 2021]	0.12	0.69	24.31	0.8434	6.115	×
	LAN [Raimundo *et al.*, 2022]	0.05	0.28	24.30	0.8224	6.146	×
	MicroISP [Ignatov *et al.*, 2023a]	**0.02**	**0.18**	23.61	0.8205	6.978	×
	PyNet-v2 [Ignatov *et al.*, 2022a]	0.98	2.11	23.96	0.8218	6.109	×
	LiteISP [Zhang *et al.*, 2021]	11.9	24	24.89	0.8577	5.801	×
	Ours (MetaISP)	0.10	0.65	**25.06**	**0.8658**	**5.576**	×
	MicroISP† [Ignatov *et al.*, 2023a]	0.02	0.18	23.87	0.8530	–	✓
	PyNet-v2† [Ignatov *et al.*, 2022a]	0.98	2.11	24.72	0.8783	–	✓

Table 1: Quantitative results for ZRR [Ignatov *et al.*, 2020b] and MAI21 datasets [Ignatov *et al.*, 2022b]. Our MetaISP achieves superior performance across all metrics, with significantly smaller FLOPs and parameter counts. The PyNet-v2 and MicroISP models denoted with † for the MAI21 dataset utilized extra data, comprising 99K pairs of training images (not released). The other models, in contrast, used 23K pairs of images. AWNet presents two versions: AWNet$_3$ and AWNet$_3$, which process 3-channel demosaicked images and 4-channel raw images as inputs respectively. ΔE is widely used to measure changes in visual perception between two colors, we report the results of ΔE 2000 [Sharma *et al.*, 2005].

LiteISP model, particularly in processing high-resolution images. Additionally, the analysis reveals that the computational complexity of the MetaISP model is linearly correlated with the input size. These efficiency assessments were conducted using a single Titan X GPU.

MAI21 Dataset. An extensive evaluation on the MAI21 dataset affirmed MetaISP's superior performance and computational efficiency compared to state-of-the-art models (see Table 1). Remarkably, MetaISP outshone models like MicroISP and PyNet-V2 in the majority of evaluation metrics, despite these models leveraging additional data. The formidable efficiency of MetaISP becomes evident when considering that it utilizes a scant 1.1% of LiteISP's parameters and 2.4% of its computational burden (in FLOPs), while still managing to outstrip LiteISP in performance. In terms of image quality, sharpness, and color restoration, MetaISP's output distinctly excels, a fact clearly illustrated in Figure 6. It is particularly noteworthy that our model significantly outpaces other models in the aspect of color information restoration.

Resolution	Method	FLOPs↓	Memory↓	Latency↓	FPS↑
224×224	LiteISP	18.3G	238MB	30ms	34
	MetaISP	**1.9G**	**155MB**	**21ms**	**47**
448×448	LiteISP	73.2G	823MB	67ms	15
	MetaISP	**7.8G**	**596MB**	**30ms**	**33**
672×672	LiteISP	164.7G	1801MB	128ms	8
	MetaISP	**17.6G**	**1345MB**	**56ms**	**18**
896×896	LiteISP	292.8G	3185MB	216ms	5
	MetaISP	**31.2G**	**2368MB**	**95ms**	**11**
1120×1120	LiteISP	457.6G	4959MB	323ms	3
	MetaISP	**48.9G**	**3712MB**	**146ms**	**7**

Table 2: Actual efficiency comparison under different resolutions.

4.3 Ablation Studies

In this section, we conduct extensive experiments to measure the contributions of our proposed blocks. Experiments are performed on the aligned ZRR dataset, and models are trained on image patches of 448×448 for 85 epochs.

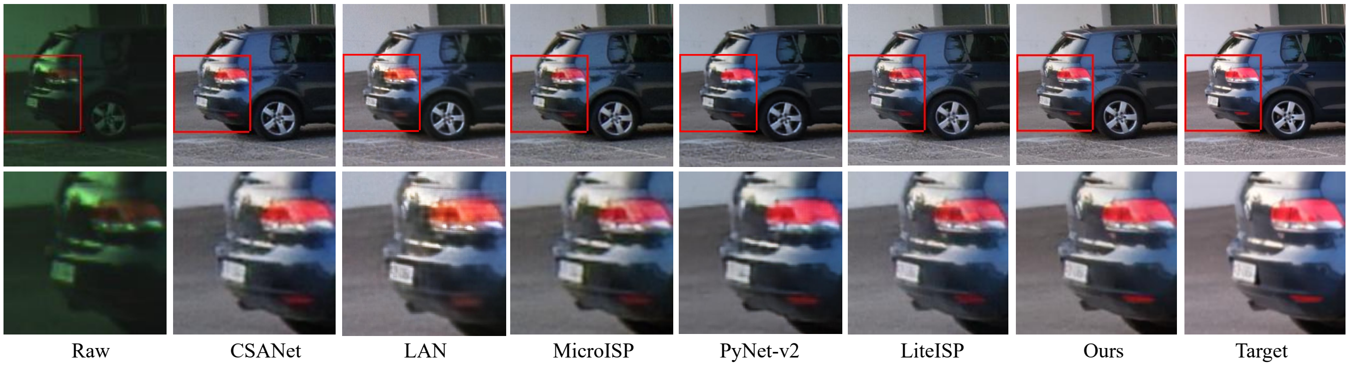


Figure 6: Visual comparison results on the MAI21 dataset [Ignatov *et al.*, 2022b].MetaISP demonstrates significant advantages in both detail preservation and global color fidelity. Best viewed with Zoom.

Effectiveness of the Four Proposed Blocks

To evaluate each individual component in MetaISP, we conducted additional experiments as shown in Table 3. The baseline is the MetaISP trained with all four components. When replacing MFEB with a commonly used 3×3 convolution layer, the PSNR decreases by 0.11 dB. The inclusion of MRB helps mitigate the overfitting issue in the model, as evidenced by a PSNR drop of 0.29 dB when removing it. In addition to enhancing the quality of output images, MFEB and MRB contribute only marginally to the computational complexity of the model. MCCB and MSCB serve as fundamental components within the U-shaped encoder-decoder architecture of MetaISP. Removing MCCB results in a PSNR drop of approximately 0.5 dB, while removing MSCB leads to a larger drop of 0.63 dB.

Various Numbers of Basis Matrices and Vectors

To determine the number N of basis matrices and vectors, we evaluate the performance of our model by setting $N = \{4, 8, 10, 16, 32\}$ as shown in Table 3. We have chosen to set N as 10 in order to strike a balance between efficiency and performance in our model. Further increasing the value of N does not result in significant performance improvements and may potentially lead to overfitting issues.

Design of MFEB

The results presented in Table 3 demonstrate that the removal of global feature extraction kernels or local feature extraction kernels degrades the performance of the baseline model. This observation suggests that both the global branches and local branches within MFEB contribute to the flexible shallow feature extraction process, thereby enhancing the overall performance of our model. Additionally, the utilization of these pre-defined kernels has a negligible impact on the number of parameters and FLOPs of the model.

Superiority to other existing blocks

For further comparative analysis, we substituted our proposed MSCB and MCCB with the widely-utilized CNN-based RCAB [Hu *et al.*, 2018] and the Self-attention-based MDTA [Zamir *et al.*, 2022]. As depicted in Table 4, our blocks demonstrate comparable efficiency to CNN-based approaches while outperforming the self-attention mechanism in terms of overall performance.

Ablation study on 4 main components of MetaISP

MFEB	MCCB	MSCB	MRB	Params(M)↓	FLOPs(G)↓	PSNR(dB)↑	SSIM↑
✓	✓	✓	✓	1.10	7.81	23.80	0.8758
×	✓	✓	✓	1.10	7.64	23.69	0.8716
✓	×	✓	✓	0.57	5.39	23.30	0.8608
✓	✓	×	✓	0.89	5.93	23.17	0.8564
✓	✓	✓	×	1.09	7.81	23.51	0.8697
×	✓	✓	×	1.09	7.64	23.66	0.8715

Ablation study on design of MFEB

Conv3	GELU	Global	Local	Params(M)↓	FLOPs(G)↓	PSNR(dB)↑	SSIM↑
✓	✓	✓	✓	1.10	7.81	23.80	0.8758
✓	×	✓	✓	1.10	7.81	23.75	0.8707
✓	✓	×	✓	1.10	7.76	23.73	0.8692
✓	✓	✓	×	1.10	7.74	23.63	0.8660
✓	✓	×	×	1.10	7.64	23.69	0.8716

Ablation study on MCCB and MSCB (The number of basis matrices and vectors)

Num	Params(M)↓	FLOPs(G)↓	PSNR(dB)↑	SSIM↑
4	0.90	7.81	23.74	0.8703
8	1.04	7.81	23.75	0.8743
10	1.10	7.81	23.80	0.8758
16	1.31	7.81	23.79	0.8757
32	1.87	7.81	23.68	0.8698

Table 3: Performance of the proposed MetaISP framework under different module configurations.

Block	Params↓	FLOPs↓	Latency↓	PSNR↑	SSIM↑
RCAB	1.02M	11.5G	20ms	23.02dB	0.8501
MDTA	0.9M	11.2G	57.2ms	23.58dB	0.8657
Ours	1.1M	7.8G	29.7ms	23.80dB	0.8758

Table 4: Replace our blocks with other existing attention blocks

5 Conclusion

In this paper, we propose a novel learnable ISP model called MetaISP. Our method aims to learn accurate mappings between RAW and sRGB under diverse environmental conditions, while significantly reducing computational complexity. This objective is accomplished by employing adaptive model parameters and architecture adjustments according to the characteristics of input images. Experiments on two large-scale public datasets demonstrate that our model outperforms the state-of-the-art methods while having substantially fewer parameters and FLOPs. Extensive ablation experiments provide compelling evidence for the effectiveness of the four main components.

Ethical Statement

This research strictly follows the highest ethical standards in image restoration. We responsibly and ethically utilized data sources, including images and algorithms, throughout the study. We have adhered to institutional and governmental guidelines for ethical research practices, ensuring transparency and reproducibility in our methodology. The research does not involve studies with human participants or animals conducted by any of the authors. In cases where external contributions were incorporated, proper credit was given to maintain academic integrity and uphold scientific community standards.

Acknowledgments

This research is supported by the Singapore Ministry of Education Academic Research Fund Tier 1 (WBS: A0009440-01-00).

Contribution Statement

Zigeng Chen and Chaowei Liu contributed equally to this work.

References

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[Chen *et al.*, 2021] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[Dai *et al.*, 2020] Linhui Dai, Xiaohong Liu, Chengqi Li, and Jun Chen. Awnet: Attentive wavelet network for image isp. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 185–201. Springer, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[Hsyu *et al.*, 2021] Ming-Chun Hsyu, Chih-Wei Liu, Chao-Hung Chen, Chao-Wei Chen, and Wen-Chia Tsai. Csanet: High speed channel spatial attention network for mobile isp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2486–2493, 2021.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[Ignatov *et al.*, 2020a] Andrey Ignatov, Radu Timofte, Zhilu Zhang, Ming Liu, Haolin Wang, Wangmeng Zuo, Jiawei Zhang, Ruimao Zhang, Zhanglin Peng, Sijie Ren, et al. 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*, pages 152–170. Springer, 2020.

[Ignatov *et al.*, 2020b] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 536–537, 2020.

[Ignatov *et al.*, 2022a] Andrey Ignatov, Grigory Malivenko, Radu Timofte, Yu Tseng, Yu-Syuan Xu, Po-Hsiang Yu, Cheng-Ming Chiang, Hsien-Kai Kuo, Min-Hung Chen, Chia-Ming Cheng, et al. Pynet-v2 mobile: Efficient on-device photo processing with neural networks. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 677–684. IEEE, 2022.

[Ignatov *et al.*, 2022b] Andrey Ignatov, Radu Timofte, Shuai Liu, Chaoyu Feng, Furui Bai, Xiaotao Wang, Lei Lei, Ziyao Yi, Yan Xiang, Zibin Liu, et al. Learned smartphone isp on mobile gpus with deep learning, mobile ai & aim 2022 challenge: Report. *arXiv preprint arXiv:2211.03885*, 2022.

[Ignatov *et al.*, 2023a] Andrey Ignatov, Anastasia Sycheva, Radu Timofte, Yu Tseng, Yu-Syuan Xu, Po-Hsiang Yu, Cheng-Ming Chiang, Hsien-Kai Kuo, Min-Hung Chen, Chia-Ming Cheng, et al. Microisp: Processing 32mp photos on mobile devices with deep learning. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 729–746. Springer, 2023.

[Ignatov *et al.*, 2023b] Andrey Ignatov, Radu Timofte, Shuai Liu, Chaoyu Feng, Furui Bai, Xiaotao Wang, Lei Lei, Ziyao Yi, Yan Xiang, Zibin Liu, et al. Learned smartphone isp on mobile gpus with deep learning, mobile ai & aim 2022 challenge: report. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 44–70. Springer, 2023.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 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*, pages 694–711. Springer, 2016.

[Kim *et al.*, 2020] Byung-Hoon Kim, Joonyoung Song, Jong Chul Ye, and JaeHyun Baek. Pynet-ca: enhanced pynet with channel attention for end-to-end mobile image signal processing. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 202–212. Springer, 2020.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Liang *et al.*, 2021a] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[Liang *et al.*, 2021b] Zhetong Liang, Jianrui Cai, Zisheng Cao, and Lei Zhang. Cameranet: A two-stage framework for effective camera isp learning. *IEEE Transactions on Image Processing*, 30:2248–2262, 2021.

[Liu *et al.*, 2018] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 773–782, 2018.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[Raimundo *et al.*, 2022] Daniel Wirzberger Raimundo, Andrey Ignatov, and Radu Timofte. Lan: Lightweight attention-based network for raw-to-rgb smartphone image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 808–816, 2022.

[Schwartz *et al.*, 2018] Eli Schwartz, Raja Giryes, and Alex M Bronstein. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE Transactions on Image Processing*, 28(2):912–923, 2018.

[Sharma *et al.*, 2005] Gaurav Sharma, Wencheng Wu, and Edul N Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 30(1):21–30, 2005.

[Sun *et al.*, 2018] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2022] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022.

[Zamir *et al.*, 2022] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.

[Zhang *et al.*, 2018] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[Zhang *et al.*, 2021] Zhilu Zhang, Haolin Wang, Ming Liu, Ruohao Wang, Jiawei Zhang, and Wangmeng Zuo. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4348–4358, 2021.