Multispectral Demosaicing via Dual Cameras

SaiKiran Tedla[*,1,2]   Junyong Lee[*,1]   Beixuan Yang[2]   Mahmoud Afifi[1]   Michael S. Brown[1,2]

[1]AI Center-Toronto, Samsung Electronics   [2]York University

{s.tedla,j.lee8,m.afifi1,michael.b1}@samsung.com   {tedlasai,byang,mbrown}@yorku.ca

Abstract

Multispectral (MS) images capture detailed scene information across a wide range of spectral bands, making them invaluable for applications requiring rich spectral data. Integrating MS imaging into multi-camera devices, such as smartphones, has the potential to enhance both spectral applications and RGB image quality. A critical step in processing MS data is demosaicing, which reconstructs color information from the mosaic MS images captured by the camera. This paper proposes a method for MS image demosaicing specifically designed for dual-camera setups where both RGB and MS cameras capture the same scene. Our approach leverages co-captured RGB images, which typically have higher spatial fidelity, to guide the demosaicing of lower-fidelity MS images. We introduce the Dual-camera RGB-MS Dataset – a large collection of paired RGB and MS mosaiced images with ground-truth demosaiced outputs – that enables training and evaluation of our method. Experimental results demonstrate that our method achieves state-of-the-art accuracy compared to existing techniques.

1. Introduction

Multispectral (MS) imaging extends beyond standard RGB imaging by capturing spectral information across multiple wavelengths, often including visible and near-infrared spectra. This enables precise analysis for applications such as agriculture, medical imaging, and environmental monitoring [2, 12, 29, 50]. MS data has also shown great potential for image enhancement [37, 43, 57, 58], making it a valuable addition to imaging pipelines.

As multi-camera systems become more common in modern smartphones, interest in integrating MS and RGB imaging has increased to leverage additional spectral data that can complement RGB images. While hyperspectral (HS) images provide denser and more contiguous spectral information than MS images, MS sensors are more practical
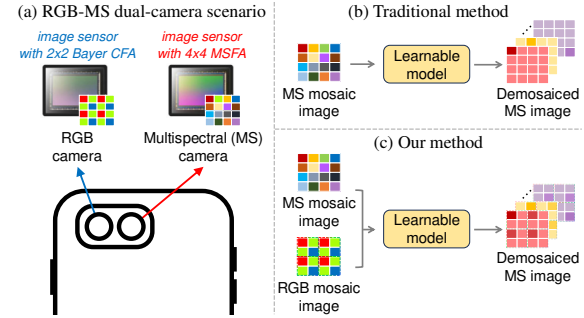

Figure 1. We propose a learning-based multispectral (MS) demosaicing method for a practical scenario (a), assuming a mobile device with a dual-camera setup, equipped with both RGB and MS cameras. Unlike traditional MS demosaicing approaches (b), our method (c) leverages the higher spatial resolution of the RGB mosaic from the RGB camera to guide and enhance the quality of the demosaiced MS image, achieving state-of-the-art results.

for mobile devices, as HS imaging typically requires expensive and time-consuming capture systems [4, 16, 42]. In turn, there has been growing research focused on incorporating MS sensors into mobile devices, demonstrating their ability to enhance performance in mobile RGB imaging tasks such as illuminant spectral estimation [18], image restoration [43], low-light enhancement [37], and tone adjustment [58]. However, most methods leverage MS imaging as a complementary prior to improving RGB-targeted tasks rather than focusing on enhancing MS image quality directly. This is primarily due to the lower fidelity typically associated with the MS imaging pipeline.

A standard RGB imaging pipeline adopts color filter arrays (CFAs) over the image sensor to capture red, green, and blue spectral bands. Each pixel records a single band, resulting in a single-channel mosaic raw image. A commonly used CFA is the Bayer array [5], which arranges the bands in a 2×2 mosaic pattern. Demosaicing algorithms then reconstruct the 3-channel RGB image by estimating the missing color values using the partial mosaic data [30].

In contrast, an MS imaging pipeline employ more complex multispectral filter arrays (MSFAs), often arranged in 4×4 mosaic patterns, to capture multiple spectral bands [3, 19, 26, 36]. While MSFA provides richer spectral in-

formation, the increased number of bands results in sparser mosaic data, making the demosaicing process considerably more challenging than in RGB demosaicing [3].

To illustrate, consider a practical example of smart-phones integrating RGB and MS cameras in a multi-camera setup (Fig. 1a). Both cameras share identical lenses and sensors in this configuration, with the RGB and MS cameras using the 2×2 Bayer CFA and 4×4 MSFA, respectively. Although both cameras capture mosaic raw images at the same resolution, MS demosaicing is inherently more complex. Meanwhile, RGB demosaicing benefits from fewer missing pixels per color channel and denser spatial data.

This observation motivates us to develop a method specifically for MS demosaicing. Our method leverages the increasing potential of integrating RGB and MS cameras within the same device [18, 28, 41]. In particular, we utilize the RGB camera image as guidance in MS demosaicing (Fig. 1c), using its higher spatial fidelity to compensate for the MS image captured with MSFAs, which trades spatial resolution for more spectral channels compared to RGB CFAs. This ensures that the reconstructed MS images preserve rich spectral information and achieve the same resolution and spatial fidelity as the RGB counterparts.

Contribution We propose an MS demosaicing method that leverages high-fidelity RGB images to address the low-fidelity nature of MS mosaic raw images (Fig. 1). We focus on a mobile setup with RGB and MS cameras mounted on a smartphone. As to the best of our knowledge, no handheld device with dual RGB and MS cameras provides access to both RGB and MS images, leading us to introduce a dual-camera RGB-MS dataset to train and validate our model. Unlike existing datasets [18, 47, 58], which lack ground-truth MS images, our dataset provides high-fidelity ground-truth MS images with detail comparable to RGB counterparts, enabling accurate evaluation of RGB-guided MS demosaicing. Training on this dataset, our model learns to leverage high-resolution RGB data, achieving state-of-the-art performance and demonstrating the potential of dual-camera systems to enhance MS image quality significantly.

2. Related work

We first define the relation of RGB, MS, hyperspectral (HS) images and how they are captured. We note the boundaries between MS and HS are not well-defined, but we utilize the definitions from [18]. First, RGB images capture spectral content that is integrated across three filters. RGB data is commonly captured on Bayer [5] sensors (2×2 CFA) by trading some spatial resolution for color information. Next, MS image capture spectral content integrated across more filters. In our case, we examine cameras with 16 filters arranged in a 4×4 CFA. Finally, HS images capture spectral content (without demosaicing) in a large number of contiguous spectral bands, offering high spectral resolution, but

are impractical for smartphones due to expensive and slow scanning setups [4, 16, 42]. Our framework is designed for real-time MS imaging, making it feasible for mobile devices and practical applications. We now discuss the most relevant works to ours which are those of HS reconstruction and RGB/MS demosaicing.

HS Image Reconstruction HS image reconstruction methods generally fall into three categories: (1) spectral super-resolution [6, 40], enhancing spectral resolution from high-resolution RGB/MS images (e.g., 3-ch RGB or 16-ch MS) to HS images; (2) spatial super-resolution [45, 56], improving the spatial resolution of HS images while preserving spectral detail; and (3) a hybrid approach [15, 25, 32, 51], using RGB/MS images to guide HS reconstruction by combining the spatial advantages of RGB/MS images with the spectral richness of HS images. Our method is quite similar to hybrid HS image reconstruction [15, 32, 51], since we leverage spatial advantages of RGB images for high-fidelity MS reconstruction.

Demosaicing RGB demosaicing, extensively studied through classical signal processing and recent learning-based methods [17, 21, 22, 33, 34, 48, 55], use CFA mosaics (typically 2×2 Bayer [5]) to leverage dense spatial information for high-quality reconstruction. In contrast, MS demosaicing, though less explored, commonly works with 4×4 MSFA mosaics and addresses challenges related to sparse spatial data [13, 14]. The limitations of both RGB and MS demosaicing motivate our framework, which combines the strengths of both approaches. Dense spatial information of RGB mosaics helps guide high-quality MS image reconstruction, despite their limited spectral content.

3. Multispectral Demosaicing

Figure 2 shows an overview of our MS demosaicing framework, specifically designed to handle both MS and RGB mosaic images captured in a dual-camera setup. Our framework takes MS and RGB mosaic images of the same scene captured with disparity and produces a demosaiced MS image while utilizing high-fidelity details of the RGB image.

We build our framework in two stages: demosaicing (Sec. 3.1) and fusion (Sec. 3.2). In the first demosaicing stage, we employ two models to independently process the MS and RGB mosaic images. In the fusion stage, a fusion module integrates the high-fidelity information from the demosaiced RGB image into the demosaiced MS image while addressing spatial disparities and spectral differences between the two images to produce the final demosaiced MS image with enhanced fidelity.

3.1. MS and RGB Demosaicing

In this stage, given the dual camera MS and RGB mosaic images, $I_{4\times4}^{MS} \in \mathbb{R}^{H\times W}$ and $I_{2\times2}^{RGB} \in \mathbb{R}^{H\times W}$, the goal is
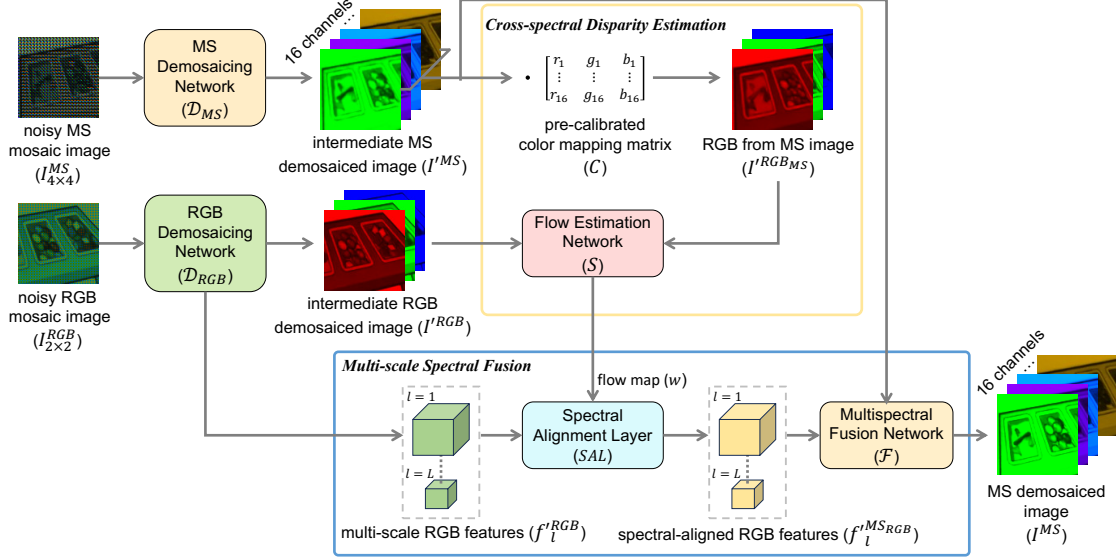
Figure 2. Our framework enhances MS demosaicing by integrating high-fidelity details from the co-captured RGB image. The framework consists of two stages: demosaicing (Sec. 3.1) and fusion (Sec. 3.2). In the demosaicing stage, the MS and RGB networks, $\mathcal{D}_{MS}$ and $\mathcal{D}_{RGB}$, reconstruct the MS and RGB images, I'^{MS} and I'^{RGB}, respectively. In the fusion stage, high-fidelity details from the RGB image are fused into the MS image, while addressing both geometric and spectral disparities. The *Cross-spectral Disparity Estimation* module computes flow map w between MS and RGB images by first transforming the MS image into RGB color space to ensure spectral compatibility during flow estimation. Then, the *Spectral Alignment Layer* (SAL) refines multi-scale RGB demosaicing features f'^{RGB}_l into $f'^{MS_{RGB}}_l$, simultaneously compensating for geometric and spectral differences to align them with the MS image. Finally, the *Multispectral Fusion Network* ($\mathcal{F}$) integrates refined RGB features $f'^{MS_{RGB}}_l$ into I'^{MS}, producing a high-fidelity MS image I^{MS}.

to reconstruct their demosaiced images $I'^{MS} \in \mathbb{R}^{H \times W \times 16}$ and $I'^{RGB} \in \mathbb{R}^{H \times W \times 3}$ using the demosaicing networks $\mathcal{D}_{MS}$ and $\mathcal{D}_{RGB}$, respectively. Here, 4×4 and 2×2 denote the mosaic pattern for MS and RGB mosaic images, respectively, and H and W represent the height and width of the images. Formally, we have:

$$I'^{MS} = \mathcal{D}_{MS}(I^{MS}_{4 \times 4}), \tag{1}$$

$$I'^{RGB} = \mathcal{D}_{RGB}(I^{RGB}_{2 \times 2}), \tag{2}$$

where we employ NAFNet [7] as backbone networks for $\mathcal{D}_{MS}$ and $\mathcal{D}_{RGB}$, selected based on their reliable performance in MS and RGB demosaicing tasks [3, 27].

Note that we perform separate demosaicing for the MS and RGB mosaic images to enable more precise and effective fusion in the subsequent stage, rather than attempting to fuse them directly. Aligning mosaic images with different patterns and spectral bands presents significant challenges, complicating the accurate fusion of MS and RGB images.

3.2. Cross-Spectral Multi-Scale Fusion

The RGB image $I'^{RGB} \in \mathbb{R}^{H \times W \times 3}$, restored from 2×2 mosaics, captures more high-frequency spatial details of a scene in contrast to the MS image $I'^{MS} \in \mathbb{R}^{H \times W \times 16}$ reconstructed from 4×4 mosaics. In this stage, our goal is to transfer the high-frequency spatial details from I'^{RGB} to I'^{MS}, generating I^{MS} with enhanced details.

Although I'^{RGB} and I'^{MS} capture the same scene, it is not straightforward to directly utilize I'^{RGB} for enhancing I'^{MS}, as the image pair is misaligned, due to the disparity introduced within the dual-camera setup, and each image contains different spectral information. To address this, we compose this stage into two modules: first, *Cross-spectral Disparity Estimation* that computes dense-correspondences between I'^{MS} and I'^{RGB}; and second, *Multi-scale Spectral Fusion* that integrates high-frequency spatial details of I'^{RGB} with fewer spectral bands into I'^{MS} with more spectral measurements to produce a final MS image I^{MS}, while compensating for the disparity between the images. In the following, we describe each module in more detail.

Cross-Spectral Disparity Estimation Computing dense correspondence between images with different spectra remains as a challenging problem [8, 31]. While some studies focus on cross-modal matching, they are limited to specific cases like RGB with Near InfraRed [23], InfraRed [9], varying illuminations [20], or MS images with limited spectral bands [47]. A recent high-resolution (HS) image super-resolution approach [25] uses stereo RGB images as guidance. For RGB image alignment, the method computes per-pixel flow between the HS and RGB images by first converting the HS image to RGB, leveraging the spectral response functions of both HS and RGB images. This process is relatively straightforward, as the HS image contains dense, contiguous spectral information that spans the RGB spectrum.

Inspired by this, we estimate cross-spectral disparity estimation between I'^{MS} and I'^{RGB} by employing a pre-calibrated color conversion matrix $C \in \mathbb{R}^{16 \times 3}$ to transform $I'^{MS} \in \mathbb{R}^{H \times W \times 16}$ into the proxy RGB $I'^{RGB_{MS}} \in \mathbb{R}^{H \times W \times 3}$, aligning it to the color space of the RGB image I'^{RGB}. Mathematically, we have:

$$I'^{RGB_{MS}} = r^{-1}\left(r\left(I'^{MS}\right) \cdot C\right), \tag{3}$$

where $r\left(I'^{MS} \in \mathbb{R}^{H \times W \times 16}\right) \to I_r'^{MS} \in \mathbb{R}^{(H \times W) \times 16}$ is a re-shaping operator and the color conversion matrix C is pre-calibrated by computing a least squares transformation between the RGB and MS color chart image pairs (refer to the supplementary material for more details).

Then, we estimate the optical flow $w \in \mathbb{R}^{H \times W \times 2}$ between $I'^{RGB_{MS}}$ and I'^{RGB} using the pre-trained flow estimation network $\mathcal{S}$ [46]. Specifically, w is obtained as:

$$w = \mathcal{S}(I'^{RGB_{MS}}, I'^{RGB}). \tag{4}$$

Note that the flow estimation network $\mathcal{S}$ first preprocesses I'^{RGB} and $I'^{RGB_{MS}}$, mapping them into the sRGB color space using corresponding camera metadata (i.e., white balance and color correction matrices) to align the images with the color space used for the flow estimation task.

Multi-Scale Spectral Fusion We now fuse the high-fidelity details of the RGB image I'^{RGB} into the MS image I'^{MS} to produce the final MS image I^{MS}, while compensating for the disparity between the two images using the estimated flow map w. We propose a multi-scale spectral fusion network $\mathcal{F}$ to address this challenge.

The network $\mathcal{F}$ takes I'^{MS} as its primary input to produce the final enhanced output I^{MS}. We introduce a *spectral alignment layer* (SAL) to incorporate details from the RGB image. The layer takes L-level multi-scale RGB feature maps $\{f_l'^{RGB}, \ l \in [1, 2, ..., L]\}$ extracted from the RGB demosaicing network and flow map w to provide refined RGB feature map $f_l'^{MS_{RGB}}$ to each level of the fusion network $\mathcal{F}$. We formally define the fusion process as:

$$f_l'^{MS_{RGB}} = SAL(f_l'^{RGB}, w), \tag{5}$$

$$I^{MS} = \mathcal{F}(I'^{MS}, \ f_{l \in [1,2,...,L]}'^{MS_{RGB}}), \tag{6}$$

where we adopt NAFNet [7] as the backbone for the network $\mathcal{F}$. Here, the refined RGB feature maps $f_l'^{MS_{RGB}}$ are aligned with the MS image and adapted for fusion with the intermediate MS features within the network $\mathcal{F}$.

SAL addresses geometric and spectral disparities simultaneously using the deformable convolution network (DCN) [11], which integrates seamlessly with the flow map w and is particularly effective at capturing spatial features through adaptive sampling patterns. SAL computes the refined RGB feature map $f_l'^{MS_{RGB}}$ as the following:

$$f_l'^{MS_{RGB}}(p) = \sum_{i \in \Omega} k(i) f_l'^{RGB}(p + p_w + i + \Delta i), \tag{7}$$

where p is the location on the output feature map $f_l'^{MS_{RGB}}$, and i enumerates the locations Ω in the deformable convolution kernel k. The optical flow offset at location p, denoted as $p_w = w(p)/2^{l-1}$, represents the MS-to-RGB image displacement downscaled by 2^{l-1}, enabling multi-scale geometric alignment of RGB to MS features. Here, Δi represents the deformable kernel offsets learned by SAL, enabling the convolution kernel k to adapt spatially and capture fine structural details in the multi-scale RGB features $f_l'^{RGB}$, such as edges and textures.

3.3. Network Training

The training process consists of two stages: demosaicing and fusion. To train our network, we use our dual-camera RGB-MS dataset, consisting of quadruplets of mosaic MS and RGB images, each paired with ground-truth demosaic MS and RGB images, denoted as $I_{4 \times 4}^{MS}$, $I_{2 \times 2}^{RGB}$, $\hat{I}^{MS}$, and $\hat{I}^{RGB}$, respectively. Sec. 4 discusses more dataset details.

MS and RGB Demosaicing We first train the MS and RGB demosaicing networks $\mathcal{D}_{MS}$ and $\mathcal{D}_{RGB}$ independently (Sec. 3.1). For $\mathcal{D}_{MS}$, we use the L2 loss between the predicted MS demosaiced image I'^{MS}, and the ground-truth MS image $\hat{I}^{MS}$. Similarly, for the network $\mathcal{D}_{RGB}$, we apply the L2 loss between the predicted RGB image I'^{RGB}, and the ground-truth RGB image $\hat{I}^{RGB}$:

$$\mathcal{L}_{MS} = \|I'^{MS} - \hat{I}^{MS}\|_2, \tag{8}$$

$$\mathcal{L}_{RGB} = \|I'^{RGB} - \hat{I}^{RGB}\|_2. \tag{9}$$

Cross-Spectral Fusion In this stage, we train the multi-scale spectral fusion module (Sec. 3.2), which comprises the spectral alignment layer (SAL) and the fusion network $\mathcal{F}$ to incorporate high-fidelity details from I'^{RGB} into the demosaiced MS image I'^{MS}, producing an enhanced MS image I^{MS}. The MS and RGB demosaicing networks and the optical flow estimation network, $\mathcal{S}$, remain fixed during this stage. We apply an L2 loss between the final MS image I^{MS} and the ground-truth MS image $\hat{I}^{MS}$:

$$\mathcal{L}_{fusion} = \|I^{MS} - \hat{I}^{MS}\|_2. \tag{10}$$

4. Dual-Camera RGB-MS Dataset

To train and validate our network, we introduce a dataset containing quadruplets of mosaiced RGB and MS images, each paired with ground-truth demosaiced RGB and MS images (Figs. 3b-e). The ground-truth demosaiced images are high-quality captures (details of capture in Section 4.1) obtained using our imaging system that simulates an asymmetric dual-camera setup (Fig. 4). The system simulates an RGB camera capturing a scene in 3 RGB channels and an MS camera capturing the same scene in 16 multispectral channels, with a spatial disparity between them. We synthesize the mosaic images by converting the demosaiced images into 1-channel mosaics. The dataset comprises 502

(a) Scenes staged in the illumination box for constructing our dataset (b) $I_{2\times2}^{RGB}$ (c) $\hat{I}^{RGB}$ (d) $I_{4\times4}^{MS}$ (e) $\hat{I}^{MS}$

Figure 3. Representative examples from our dataset: (a) 28 scenes staged in the illumination box and (b–e) samples of quadruplets captured from scenes 7, 14, 21, and 28. Each quadruplet consists of (b) a 1-channel RGB mosaic image $I_{2\times2}^{RGB}$, (c) the corresponding 3-channel RGB demosaiced ground-truth $\hat{I}^{RGB}$, (d) a 1-channel MS mosaic image $I_{4\times4}^{MS}$, and (e) its corresponding 16-channel MS demosaiced ground-truth $\hat{I}^{MS}$. Note that the images within each quadruplet share the same spatial resolution. We also visualize the mosaic patterns (zoomed-in red cropped box) and the disparity between the RGB and MS mosaic images (white dashed line) in the first row.

Figure 4. Dual-camera MS-RGB capturing system.

quadruplets across 28 challenging scenes (Fig. 3a) with high textures and detailed features. The dataset has training, validation, and test sets, containing 352, 47, and 103 image quadruplets captured from 20, 2, and 6 scenes, respectively, where each image is in camera raw space at the resolution of 1440×2160 pixels. Additional details about capture setup are given in the supplementary material.

Existing MS datasets [18, 58] are not well-suited for our task due to the absence of high-quality, ground-truth demosaiced images alongside the MS data. Additionally, these datasets often feature very low-resolution MS and RGB images in sRGB space already processed by camera pipeline [58], or paired RGB and MS images with minimal scene overlap due to extremely large disparities [18]. Our method, designed for handheld devices like smartphones with minimal multi-camera disparity, led us to collect a dataset with realistic disparity and accurate ground-truth de-

mosaic images. It includes both RGB and synthetic MS images in raw space, with no processing from the camera pipeline, making it ideal for training and evaluating our approach in the early stages of the onboard camera pipeline.

4.1. Data Gathering Pipeline

To collect the dataset, we build an imaging system consisting of a camera mounted on a linear stage actuator and a controllable illumination box (Fig. 4), that allows us to simulate an asymmetric dual-camera setup, where MS and RGB cameras have a constant relative baseline.

The linear stage actuator moves the camera to two adjustable positions, one for MS and the other for RGB capture. We utilize an Arduino/Genuino microcontroller to precisely control the camera movement between positions, ensuring a constant relative baseline between MS and RGB captures. In practice, we set the baseline to 1 cm. The controllable illumination box (Telelumen Octa Light Player) simulates multispectral data capturing, featuring configurable light sources that distribute evenly throughout the scene within the box. Capturing occurs in a lab setting in a dark room to ensure that the box is the sole lighting source.

We use a Sony Alpha 1 camera as the capturing device. For each capture, the system generates a pair of 16-channel MS and 3-channel RGB demosaic images using the pixel shift mode featured in the camera, which shifts the sensor during capture to enable sensor-level demosaicing. Each image is initially captured at a resolution of 5760×8640 pixels. To mitigate noise introduced by the small pixel size of the sensor, we downsample these images to a resolution of 1440×2160 pixels to generate ground-truth demosaiced images. We create our 1-channel MS and RGB input mosaic images by applying a 4×4 MSFA pattern and 2×2 Bayer CFA pattern to the MS and RGB ground-truth demosaiced images, respectively. Additionally, we simulate noise on the mosaic images using a Poisson-Gaussian noise model [38], calibrated at ISO 400. For this calibration, we use 90 im-

ages of a color chart (30 images taken at three different exposures) captured by the camera.

Like most consumer cameras, the camera in our system uses a CFA that captures the RGB spectrum, making RGB image acquisition straightforward. We obtain RGB images by configuring the light sources within the controllable illumination box to simulate the CIE D65 daylight illuminant. For MS image acquisition, we simulate an MSFA by capturing multiple RGB images of a scene under varying light sources. The following provides further background for simulating MS capture using the illumination box.

Multispectral Image Acquisition To begin, we consider the image formation process in our setup under a uniform light source across the scene. Formally, the color information of a mosaic image I at location x can be described as:

$$I(x) = \int_\gamma \underbrace{S(y)C_{rgb}(x,y)}_{CRF_{RGB}} \underbrace{L(y)R(x,y)}_{scene\ irradiance} \, dy + z, \quad (11)$$

where CRF_{RGB} represents the RGB camera response function (composed of the sensor's spectral sensitivity $S(y)$ and the RGB CFA response function $C_{rgb}(x,y)$), $L(y)$ is the spectral power distribution (SPD) of light, and $R(x,y)$ is the scene reflectance. z denotes unwanted noise, typically characterized by signal-dependent and additive components [1]. The integral over the visible range γ at wavelength y provides the color information in the mosaic raw image.

Next, let us combine the camera response function with the light SPD emitted by the illumination box in narrow spectral bands, which mimic the spectral filters used in MS systems. We assume that the scene is lit by a uniform, broadband, and neutral "virtual" light source with an SPD denoted as $J(\cdot)$, which spans all wavelengths equally. Eq. (12) can then be rewritten as:

$$I(x) = \int_\gamma \underbrace{S(y)C_{rgb}(x,y)L(y)}_{CRF_{MS}} \underbrace{J(y)R(x,y)}_{simulated\ scene\ irradiance} \, dy + z, \quad (12)$$

where CRF_{MS} represents our virtual MS camera response function, and the scene irradiance is now simulated by the scene reflectance under the assumption of a constant, uniform, broadband, neutral virtual light source.

Since our capturing system uses the illumination box as the only physical light source in the scene, the box allows us to control the SPD of the light. This capability enables us to simulate the MS mosaic image by capturing the scene with the RGB CFA, varying the SPD of the light within the illumination box, and performing multiple captures to simulate the response function of an MS camera.

The box provides seven primary wavelengths, ranging from 380 nm to 760 nm, which can be combined in various ways to create customizable light sources. For MS image acquisition, the system captures the same scene seven times, each under a different wavelength combination, resulting in a 21-channel MS image (7 wavelength combinations × 3

RGB channels). This is then reduced to a 16-channel MS image by discarding the 5 spectral channels with the least information. Detailed discussions and experiments are included in the supplementary material.

5. Experiments

We train our model using the proposed dual-camera MS-RGB dataset (Sec. 4). During training, we use the Adam optimizer [24] with a learning rate of 1×10^{-3}. Following the two-stage training strategy (Sec. 3.3), we train the model for 200k iterations in both the demosaicing and fusion stages. For each iteration, we randomly sample batches of 8 quadruplets (mosaiced MS and RGB images, along with their corresponding demosaiced outputs) from the training set and cropped into 256×256 patches. For the quantitative evaluation of MS restoration quality, we measure the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [49], and Spectral Angle Mapper (SAM) [53] .

5.1. Ablation Study

We conduct ablation studies focusing on the impact of the proposed cross-spectral multi-scale fusion module (Sec. 3.2) in incorporating RGB guidance for reconstructing the final MS demosaiced image. We compare the baseline MS demosaicing network $\mathcal{D}_{MS}$ (Eq. (1)), with NAFNet [7] as the backbone, and three variants that utilize the RGB demosaicing network $\mathcal{D}_{RGB}$ (Eq. (2)) to provide RGB guidance to the multispectral fusion network $\mathcal{F}$. The variants employ RGB guidance through image-based or feature-based methods. In the image-based guidance, the RGB demosaicked image I'^{RGB} is used directly, with or without alignment to the MS image. Each of these is concatenated with the MS demosaicked image I'^{MS} and fed into the fusion network $\mathcal{F}$. The feature-based guidance leverages configurations of the proposed spectral alignment layer (SAL).

Table 1 presents quantitative results. The baseline model performs the worst due to the absence of RGB guidance (first row). Providing RGB guidance by concatenating I'^{RGB} with the MS image I'^{MS} yields slight improvements (second row), while flow-based alignment further enhances performance by addressing geometric misalignment (third row). Incorporating multi-scale RGB features refined by the proposed SAL, which adapts deformable convolution [11], achieves the best performance (last row) and demonstrates its effectiveness in utilizing RGB features to facilitate cross-spectral fusion with multi-scale features (see supplementary for a more detailed analysis of SAL).

5.2. Comparison on MS Demosaicing

In this section, we evaluate the proposed method in two key asymmetric dual-camera scenarios. The first scenario focuses on asymmetry in the CFA patterns of the RGB and

$\mathcal{D}_{MS}$ (Eq. (1))	$\mathcal{D}_{RGB}$ (Eq. (2))	RGB guidance for fusion network $\mathcal{F}$ (Eq. (6))			PSNR↑	SSIM↑	SAM↓	Params (MB)	MACs[1] (T)
		I'^{RGB}	$I'^{RGB}(p+w(p))$	SAL					
✓					40.89	0.9766	2.604	111.24	0.78
✓	✓	✓			40.90	0.9767	2.597	124.17	1.91
✓	✓		✓		41.75	0.9808	2.520	124.17	1.91
✓	✓			✓	**41.92**	**0.9811**	**2.422**	130.03	2.53

Table 1. Ablation study evaluating the effect of RGB guidance strategies of the proposed MS demosaicing framework. We examine direct usage of the RGB demosaic I'^{RGB}, the warped image $I'^{RGB}(p+w(p))$ aligned to the MS image using the flow w (Eq. (4)), and feature-based guidance with the proposed spectral alignment layer (SAL, Eqs. (5) and (7)).

Input	Model	PSNR↑	SSIM↑	SAM↓	Params (MB)	MACs[1] (T)
$I^{MS}_{4\times4}$	SSMT [13]	33.97	0.932	6.723	9.76	26.01
	MCAN [14]	39.02	0.966	3.947	5.24	0.91
	MCAN-L	40.36	0.974	2.996	53.69	7.80
	NAFNet [7]	40.89	0.977	2.604	111.25	0.78
	NAFNet-L	40.68	0.976	2.654	158.58	1.51
	Restormer [54]	40.61	0.977	2.643	99.70	6.72
	Restormer-L	40.58	0.976	2.660	148.71	11.82
$I^{MS}_{4\times4}$ & $I^{RGB}_{2\times2}$	DCT [32]	38.44	0.962	4.748	31.91	15.22
	HSIFN [25]	36.50	0.963	3.198	90.21	18.79
	MCAN+Ours	**41.85**	**0.981**	2.572	24.03	2.66
	NAFNet+Ours	**41.92**	**0.981**	**2.422**	130.03	2.53
	Restormer+Ours	41.48	0.981	**2.474**	118.49	8.47

Table 2. Quantitative comparison for MS demosaicing.

Input	Model	PSNR↑	SSIM↑	SAM↓	Params (MB)	MACs[2] (T)
$I^{MS\downarrow4}_{4\times4}$	SSMT [13]	32.08	0.902	6.651	12.71	26.92
	MCAN [14]	31.97	0.900	5.755	5.43	0.08
	MCAN-L	32.87	0.916	4.884	53.88	0.51
	NAFSR [10]	32.98	0.917	4.736	59.19	2.66
	NAFSR-L	32.95	0.919	4.439	100.93	4.54
	Restormer [54]	32.34	0.908	5.166	99.99	0.93
	Restormer-L	32.45	0.911	4.882	149.00	1.26
$I^{MS\downarrow4}_{4\times4}$ & $I^{RGB}_{2\times2}$	DCT [32]	31.12	0.893	6.613	42.21	15.78
	HSIFN [25]	29.92	0.866	6.982	90.35	6.61
	MCAN+Ours	**37.55**	**0.962**	**3.629**	24.22	1.82
	NAFSR+ Ours	**37.67**	**0.964**	**3.567**	77.98	4.41
	Restormer+Ours	36.70	0.960	3.722	118.77	2.68

Table 3. Quantitative comparison for 4× MS demosaicing.

MS sensors. The second introduces an additional asymmetry where the MS sensor has a lower resolution than the RGB sensor, reflecting a practical consideration where a smaller MS sensor is used in the dual-camera setup.

We compare our RGB-guided MS demosaicing framework with previous image restoration methods, including hyperspectral image restoration and multispectral demosaicing methods: SSMT [13], MCAN [14], DCT [32], HSIFN [25], NAFNet [7, 10], and Restormer [54]. Among these, SSMT and MCAN are MS mosaic-to-MS reconstruction methods. NAFNet and Restormer are general-purpose image restoration models with state-of-the-art performance in tasks such as demosaicing [27, 52] and super-resolution [39, 59]. DCT and HSIFN are hybrid MS-to-HS reconstruction methods that use RGB as a guidance. All baselines are retrained from scratch on our dataset using the same training configuration as our model and all are dtrained with the same number of iterations for fairness.

For comparison, we categorize these methods into two types based on their input: a single MS mosaic ("$I^{MS}_{4\times4}$"), and a MS mosaic with an auxiliary RGB mosaic images ("$I^{MS}_{4\times4}$ & $I^{RGB}_{2\times2}$"). We adapt each model to handle the appropriate mosaiced inputs and generate demosaiced MS images. Specifically, we adapt SSMT, MCAN, NAFNet, and Restormer to work with mosaiced MS images. DCT and HSIFN are modified to accept mosaiced MS images as input, while using mosaiced RGB as an auxiliary guidance. Additional details are provided in the supplementary.

In comparison, our framework is validated in a plug-and-play manner by integrating MCAN, NAFNet, and Restormer into a multispectral (MS) demosaicing network D_{MS}, on top of the RGB demosaicing network D_{RGB} (Eq. (2)) and the cross-spectral fusion module (Sec. 3.2). To ensure fairness, we account for the increased model complexity introduced by our approach by also evaluating baseline models with increased capacity, denoted as "-L".

Scenario 1: Asymmetric CFA Pattern In this scenario, we examine MS demosaicing in an asymmetric dual-camera configuration where the MS and RGB sensors share the same spatial resolution while using CFAs with different mosaic patterns: a 4×4 MFSA for the MS sensor and a 2×2 Bayer pattern for the RGB sensor.

For comparison, the baseline methods are trained using L2 loss between the predicted demosaiced MS image and ground-truth MS images (i.e., Eq. (8)), while our models are trained using the complete pipeline described in Sec. 3.3.

Table 2 provides quantitative results. NAFNet performs best among the baseline models, followed by Restormer, MCAN, and SSMT. For capacity-increased baselines, while MCAN-L shows improvements, NAFNet-L and Restormer-L perform worse than their original versions, likely due to overfitting when using only MS mosaic input. By incorporating our proposed method (rows marked "+ Ours"), all models show significant improvements in MS restoration quality, demonstrating the adaptability of our approach across varying architectures. The qualitative results in Fig. 5 show how our method induces a baseline model to produce better details and structures. The supplementary materials provide additional examples to confirm the bene-

[1]Measured on 1440×2160 MS and RGB mosaic images.
[2]Measured on 360×540 MS and 1440×2160 RGB mosaic images.

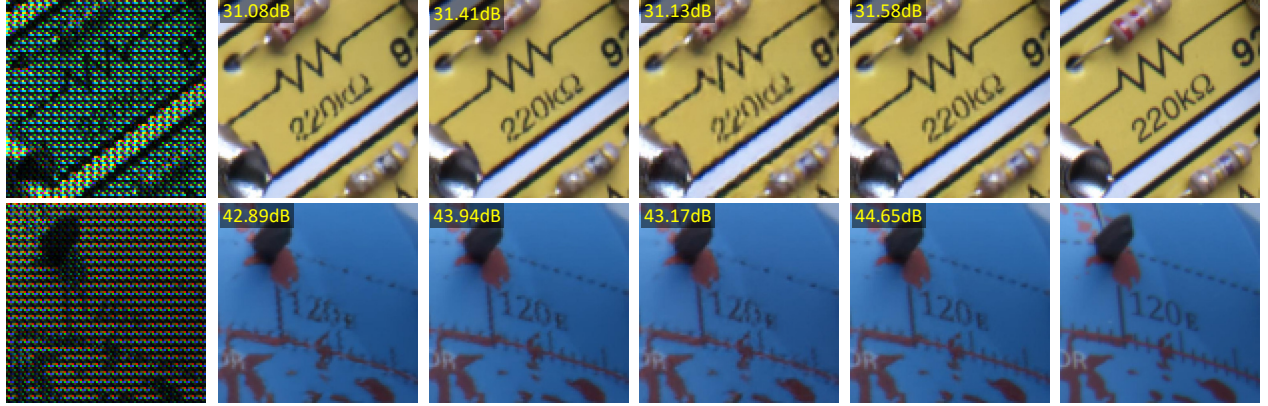| (a) MS mosaic input | (b) Restormer | (c) Restormer + Ours | (d) NAFNet | (e) NAFNet + Ours | (f) GT |

Figure 5. Qualitative comparison of MS demosaicing results for a dual-camera scenario where we consider MS and RGB sensors with the same spatial resolution but using asymmetric CFAs. Following [14, 35], we visualize the predicted MS demosaics by converting them to the sRGB color space leveraging the color conversion matrix C (Eq. (3)) and camera metadata, using CIE D65 as the reference white point.

fits of our approach in asymmetric CFA pattern scenarios.

Scenario 2: Asymmetric Sensor Resolution This scenario extends Scenario 1 by introducing an additional asymmetry in sensor resolution, where the MS and RGB sensor captures low- and high-resolution mosaics, respectively. In this scenario, we aim to reconstruct low-resolution MS mosaic images into high-resolution MS demosaiced images while preserving spectral fidelity.

For comparison, we adapt the baseline methods that takes MS mosaic image as an input to produce MS demosaic images at the desired spatial resolution. Specifically, we modify SSMT [13], DCT [32], HSIFN [25], MCAN [14], Restormer [54] by appending upsampling layers consisting of multiple convolutional and pixel-shuffle layers [44]. We employ NAFSR [10], a variant of NAFNet designed for the super-resolution task (refer supplement for details). The training follows the same procedure as in Scenario 1, except that we synthesize low-resolution MS mosaic input images by downsampling the ground-truth MS demosaic images, followed by mosaicing to simulate data captured by a smaller MS sensor. For downsampling, we use strided box filtering at a target scaling factor of 4.

Table 3 presents the quantitative results. Among the baselines, NAFSR-L achieves the best MS reconstruction performance, followed by Restormer-L and MCAN-L. Integrating our approach (rows marked "+ Ours") significantly enhances performance across all baselines, particularly in spectral accuracy, as SAM scores reflect. Unlike Scenario 1, where capacity increase led to overfitting for NAFNet and Restormer, ambiguity in dealing with low-resolution MS mosaic images allows these models to benefit from additional capacity. Figure 6 provides qualitative results. Our approach significantly improves reconstructing high-fidelity details and accurate structures, demonstrating the advantages of the proposed dual-camera scenario in enhancing MS images by leveraging high-quality RGB mosaics.

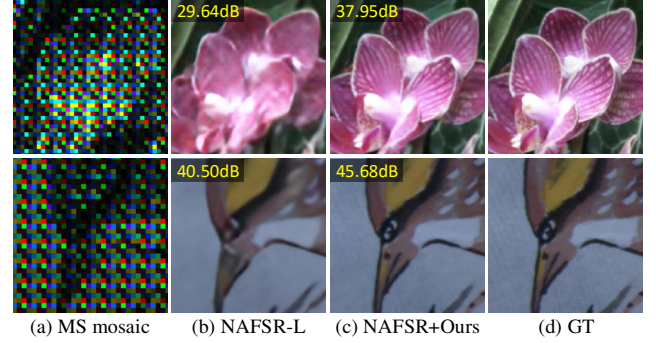| (a) MS mosaic | (b) NAFSR-L | (c) NAFSR+Ours | (d) GT |

Figure 6. Qualitative comparison of $4\times$ MS demosaicing results for a dual-camera configuration with asymmetric sensor resolution. The results are in the sRGB color space as described in Fig. 5.

6. Conclusion

This paper proposes an MS demosaicing framework leveraging high-fidelity RGB guidance from a dual-camera MS-RGB setup to enhance MS image quality. Our framework integrates a cross-spectral fusion module to address geometric misalignment and spectral disparities, effectively combining RGB and MS information. We also introduce a dual MS-RGB dataset with high-fidelity ground-truth MS images, enabling accurate evaluation of RGB-guided MS demosaicing. Experiments demonstrate the state-of-the-art MS restoration quality of our approach, highlighting the potential of dual-camera systems to advance MS imaging.

Limitation Our method enhances MS image restoration using high-fidelity RGB guidance but relies on dual RGB-MS camera setups, which are not yet widely adopted in commercial devices. Its effectiveness also depends on RGB image quality and precise cross-spectral alignment, as misalignment or noise can impact MS reconstruction. Future work could focus on reducing computational overhead to improve feasibility for resource-constrained devices.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018. 6

[2] Boaz Arad, Radu Timofte, Rony Yahel, Nimrod Morag, Amir Bernat, Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, et al. NTIRE 2022 spectral recovery challenge and data set. In *CVPRW*, 2022. 1

[3] Boaz Arad, Radu Timofte, Rony Yahel, Nimrod Morag, Amir Bernat, Yaqi Wu, Xun Wu, Zhihao Fan, Chenjie Xia, Feng Zhang, et al. NTIRE 2022 spectral demosaicing challenge and data set. In *CVPRW*, 2022. 1, 2, 3

[4] R. W. Basedow, D. C. Carmer, and M. E. Anderson. Hydice system: Implementation and performance. In *SPIE's Symposium on OE/Aerospace Sensing and Dual Use Photonics*, 1995. 1, 2

[5] Bryce E. Bayer. Color imaging array, 1976. US3971065A. 1, 2

[6] Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte, and Luc Van Gool. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *CVPRW*, 2022. 2

[7] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 3, 4, 6, 7

[8] Zhuoyuan Chen, Xun Sun, Liang Wang, Yinan Yu, and Chang Huang. A deep visual correspondence embedding model for stereo matching costs. In *ICCV*, 2015. 3

[9] Walon W. Chiu, Ulf Blanke, and Mario Fritz. Improving the kinect by cross-modal stereo. In *BMVC*, 2011. 3

[10] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: Stereo image super-resolution using nafnet. In *CVPRW*, 2022. 7, 8

[11] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 4, 6

[12] Martin De Biasio, Thomas Arnold, Raimund Leitner, Gerald McGunnigle, and Richard Meester. UAV-based environmental monitoring using multi-spectral imaging. In *Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications VII*, pages 331–337, 2010. 1

[13] Le Dong, Mengzu Liu, Tengteng Tang, Tao Huang, Jie Lin, Weisheng Dong, and Guangming Shi. Spatial-spectral mixing transformer with hybrid image prior for multispectral image demosaicing. *IEEE Journal of Selected Topics in Signal Processing*, 2024. 2, 7, 8

[14] Kai Feng, Yongqiang Zhao, Jonathan Cheung-Wai Chan, Seong G Kong, Xun Zhang, and Binglu Wang. Mosaic convolution-attention network for demosaicing multispectral filter array images. *IEEE TCI*, 7:864–878, 2021. 2, 7, 8

[15] Ying Fu, Tao Zhang, Yinqiang Zheng, Debing Zhang, and Hua Huang. Hyperspectral image super-resolution with optimized rgb guidance. In *CVPR*, 2019. 2

[16] Liang Gao, Robert T. Kester, Nathan Hagen, and Tomasz S. Tkaczyk. Snapshot image mapping spectrometer (ims) with high sampling density for hyperspectral microscopy. *Opt. Express*, 18(14):14330–14344, 2010. 1, 2

[17] Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand. Deep joint demosaicking and denoising. *ACM ToG*, 35(6):1–12, 2016. 2

[18] Ortal Glatt, Yotam Ater, Woo-Shik Kim, Shira Werman, Oded Berby, Yael Zini, Shay Zelinger, Sangyoon Lee, Heejin Choi, and Evgeny Soloveichik. Beyond RGB: A real world dataset for multispectral imaging in mobile devices. In *WACV*, 2024. 1, 2, 5

[19] Robin Hahn, Freya-Elin Hämmerling, Tobias Haist, David Fleischle, Oliver Schwanke, Otto Hauler, Karsten Rebner, Marc Brecht, and Wolfgang Osten. Detailed characterization of a mosaic based hyperspectral snapshot imager. *Optical Engineering*, 59(12):125102–125102, 2020. 1

[20] Yong Seok Heo, Kyoung Mu Lee, and Sang Uk Lee. Joint depth map and color consistency estimation for stereo images with different illuminations and cameras. *IEEE TPAMI*, 35(5):1094–1105, 2012. 3

[21] Keigo Hirakawa and Thomas W Parks. Adaptive homogeneity-directed demosaicing algorithm. *IEEE TIP*, 14(3):360–369, 2005. 2

[22] Irina Kim, Seongwook Song, Soonkeun Chang, Sukhwan Lim, and Kai Guo. Deep image demosaicing for submicron image sensors. *Electronic Imaging*, 32:1–12, 2019. 2

[23] Seungryong Kim, Dongbo Min, Stephen Lin, and Kwanghoon Sohn. Deep self-correlation descriptor for dense cross-modal correspondence. In *ECCV*, 2016. 3

[24] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[25] Zeqiang Lai, Ying Fu, and Jun Zhang. Hyperspectral image super resolution with real unaligned rgb guidance. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):2999–3011, 2025. 2, 3, 7, 8

[26] Pierre-Jean Lapray, Xingbo Wang, Jean-Baptiste Thomas, and Pierre Gouton. Multispectral filter arrays: Recent advances and practical implementation. *Sensors*, 14(11):21626–21659, 2014. 1

[27] Haechang Lee, Dongwon Park, Wongi Jeong, Kijeong Kim, Hyunwoo Je, Dongil Ryu, and Se Young Chun. Efficient unified demosaicing for bayer and non-bayer patterned image sensors. In *ICCV*, 2023. 3, 7

[28] Jaesoong Lee, Yeonsang Park, Hyochul Kim, Young-Zoon Yoon, Woong Ko, Kideock Bae, Jeong-Yub Lee, Hyuck Choo, and Young-Geun Roh. Compact meta-spectral image sensor for mobile applications. *Nanophotonics*, 11(11):2563–2569, 2022. 2

[29] Richard M. Levenson and James R. Mansfield. Multispectral imaging in biology and medicine: Slices of life. *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, 69(8):748–758, 2006. 1

[30] Xin Li, Bahadir Gunturk, and Lei Zhang. Image demosaicing: A systematic survey. In *VCIP*, 2008. 1

[31] Wenjie Luo, Alexander G. Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *CVPR*, 2016. 3

[32] Qing Ma, Junjun Jiang, Xianming Liu, and Jiayi Ma. Reciprocal transformer for hyperspectral and multispectral image fusion. *Information Fusion*, 104:102148, 2024. 2, 7, 8

[33] Henrique S Malvar, Li-wei He, and Ross Cutler. High-quality linear interpolation for demosaicing of bayer-patterned color images. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004. 2

[34] Daniele Menon, Stefano Andriani, and Giancarlo Calvagno. Demosaicing with directional filtering and a posteriori decision. *IEEE TIP*, 16(1):132–141, 2006. 2

[35] Sofiane Mihoubi, Olivier Losson, Benjamin Mathon, and Ludovic Macaire. Multispectral demosaicing using pseudo-panchromatic image. *IEEE Transactions on Computational Imaging*, 3(4):982–995, 2017. 8

[36] Yuri Murakami, Masahiro Yamaguchi, and Nagaaki Ohyama. Hybrid-resolution multispectral imaging using color filter array. *Optics Express*, 20(7):7173–7183, 2012. 1

[37] Muyao Niu, Zhihang Zhong, and Yinqiang Zheng. NIR-assisted video enhancement via unpaired 24-hour data. In *ICCV*, 2023. 1

[38] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *CVPR*, 2017. 5

[39] Zidian Qiu, Zongyao He, Zhihao Zhan, Zilin Pan, Xingyuan Xian, and Zhi Jin. Sc-nafssr: Perceptual-oriented stereo image super-resolution using stereo consistency guided nafssr. In *CVPRW*, 2023. 7

[40] Qiaoying Qu, Bin Pan, Xia Xu, Tao Li, and Zhenwei Shi. Unmixing guided unsupervised network for rgb spectral super-resolution. *IEEE TIP*, 32:4856–4867, 2023. 2

[41] Young-Geun Roh, Hyochul Kim, Wooshik Kim, Jaesoong Lee, Suyeon Lee, Youngho Jung, Heejin Choi, and Sangyoon Lee. Spectral image sensing in the smartphone and its applications. In *Image Sensing Technologies: Materials, Devices, Systems, and Applications X*, 2023. 2

[42] Yoav Y. Schechner and Shree K. Nayar. Generalized mosaicing: Wide field of view multispectral imaging. *IEEE TPAMI*, 24(10):1334–1348, 2002. 1, 2

[43] Xiaoyong Shen, Qiong Yan, Li Xu, Lizhuang Ma, and Jiaya Jia. Multispectral joint image restoration via optimizing a scale map. *IEEE TPAMI*, 37(12):2518–2530, 2015. 1

[44] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 8

[45] Oleksii Sidorov and Jon Yngve Hardeberg. Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution. In *ICCVW*, 2019. 2

[46] Zachary Teed and Jia Deng. Raft: Recurrent all pairs field transforms for optical flow. In *ECCV*, 2020. 4

[47] Fabio Tosi, Pierluigi Z. Ramirez, and Matteo Poggi. Rgb-multispectral matching: Dataset, learning methodology, evaluation. In *CVPR*, 2022. 2, 3

[48] Yi-Qing Wang. A multilayer neural network for image demosaicking. In *ICIP*, 2014. 2

[49] Zhou Wang, Alan Bovik, Hamid R. Sheikh, and Eero Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6

[50] Georgina Warren and Graciela Metternicht. Agricultural applications of high-resolution digital multispectral imagery. *Photogrammetric Engineering & Remote Sensing*, 71(5):595–602, 2005. 1

[51] Chanyue Wu, Dong Wang, Yunpeng Bai, Hanyu Mao, Ying Li, and Qiang Shen. Hsr-diff: Hyperspectral image super-resolution via conditional diffusion models. In *ICCV*, 2023. 2

[52] Senyan Xu, Zhijing Sun, Jiaying Zhu, Yurui Zhu, Xueyang Fu, and Zheng-Jun Zha. Demosaicformer: Coarse-to-fine demosaicing network for hybridevs camera. In *CVPRW*, 2024. 7

[53] Roberta H Yuhas, Alexander FH Goetz, , and Joe W Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, 1992. 6

[54] 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 *CVPR*, 2022. 7, 8

[55] Lei Zhang and Xiaolin Wu. Color demosaicking via directional linear minimum mean square-error estimation. *IEEE TIP*, 14(12):2167–2178, 2005. 2

[56] Mingjin Zhang, Chi Zhang, Qiming Zhang, Jie Guo, Xinbo Gao, and Jing Zhang. Essaformer: Efficient transformer for hyperspectral image super-resolution. In *ICCV*, 2023. 2

[57] Xiaopeng Zhang, Terence Sim, and Xiaoping Miao. Enhancing photographs with near infra-red images. In *CVPR*, 2008. 1

[58] Kailai Zhou, Lijing Cai, Yibo Wang, Mengya Zhang, Bihan Wen, Qiu Shen, and Xun Cao. Joint rgb-spectral decomposition model guided image enhancement in mobile photography. In *ECCV*, 2024. 1, 2, 5

[59] Qiang Zhu, Pengfei Li, and Qianhui Li. Attention retractable frequency fusion transformer for image super resolution. In *CVPRW*, 2023. 7