Supplementary Material for 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

1. Dual-camera MS-RGB Dataset

The proposed dataset provides high-quality image quadruplets consisting of mosaiced RGB and MS images alongside their respective demosaiced ground truths. These images are collected using a custom-built imaging system designed to simulate an asymmetric dual-camera setup, as described in Sec. 4.1 of the main paper.

Our system employs a Sony Alpha 1 camera with a RGB Bayer color filter array (CFA) sensor, featured with pixel-shift mode, to capture accurate ground-truth demosaiced RGB and MS images through sub-pixel shifts. The camera, mounted on a linear stage actuator, captures staged scenes in an illumination box from different positions while maintaining a fixed relative baseline between the RGB and MS captures (illustrated in Fig. 1). RGB acquisition is performed by configuring the illumination box to simulate the CIE D65 daylight illuminant, while MS acquisition is achieved by simulating a multispectral filter array (MSFA) through capturing the same scene under varying light sources within the box. After capturing the demosaiced RGB and MS images, further processing generates the mosaiced images. This involves synthesizing noise to replicate realistic sensor conditions and applying mosaic patterns using a 2×2 Bayer CFA for RGB images and a 4×4 MSFA for MS images.

1.1. 16-Band MS Image Acquisition

For capturing MS image acquisition, our imaging system simulates a total of 21 multispectral response functions by combining the CFA response functions of the camera with the spectral power distributions (SPDs) of varying light sources provided by the illumination box. The illumination box provides seven primary wavelengths, ranging from 380 nm to 760 nm, which can be combined in various ways to create customizable light sources. By leveraging the configurable illumination feature provided by the box, 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 5 spectral channels with the least information.

Fig. 2 illustrates the channel selection process for creating 16-channel MS images. Using an RGB camera and a configurable illumination box, we simulate 21 multispectral response functions (Fig. 2a). The response functions of the RGB CFA camera, denoted as $C^i_{rgb}(x,\gamma)$, define the sensitivity of the red, green, and blue channels across the visible wavelength range γ. The illumination box provides spectral power distributions (SPDs), $L^j(\gamma)$, corresponding to seven distinct wavelength bands. The RGB CFA response functions are calibrated using camSPECS V2, which captures images of several monochromatic light sources at different wavelengths and measures the output intensities for each channel. These measured values are compared against the known SPD of the light source to determine the spectral sensitivity of each CFA channel. For the SPDs of the illumination box, we use data provided by the manufacturer (Telelumen Octa Light Player).

The multispectral responses are derived by combining the RGB response functions with the SPDs of the box:

$$C^k_{ms}(x,\gamma) = C^i_{rgb}(x,\gamma)L^j(\gamma), {\scriptstyle \forall i\in[1,2,3],\, j\in[1,2,\ldots,7].} \quad (1)$$

This results in a total of 21 response functions (*i.e.*, $k \in [1,2,\ldots,21]$), representing all combinations of RGB channels and illumination SPDs.

To reduce the 21 channels to 16, we consider the area of the response functions and select filter responses that are distributed across the visible spectrum. Specifically, we compute the integral of each response function $C^k_{ms}(x,\gamma)$ over the visible range γ, which quantifies the spectral contribution of each channel. Based on these integrals, we select the top 12 channels with the largest areas. For the remaining 8 channels, which have smaller spectral contributions, we heuristically choose 4 channels to ensure coverage across different wavelengths. The final 16 response functions, highlighted in red in Fig. 2a, are mapped to the MSFA grid to create spatially multiplexed MS images (Fig. 2b-c). These MS images are then combined to construct the high-quality MS demosaiced image, which serves as a key component of the proposed dual-camera RGB-MS dataset.

Code and data are available at https://ms-demosaic.github.io/
*These authors contributed equally to this work.

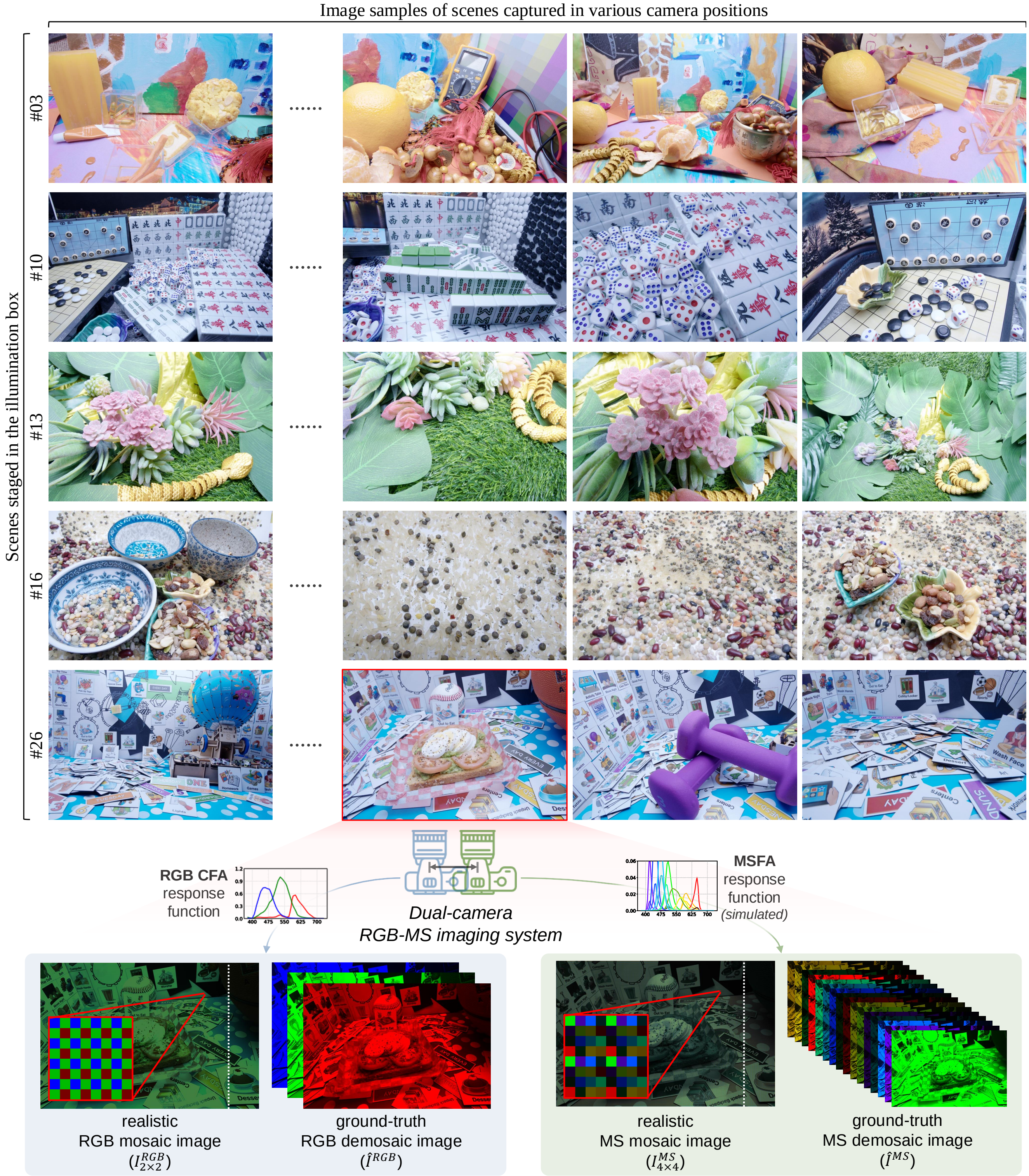
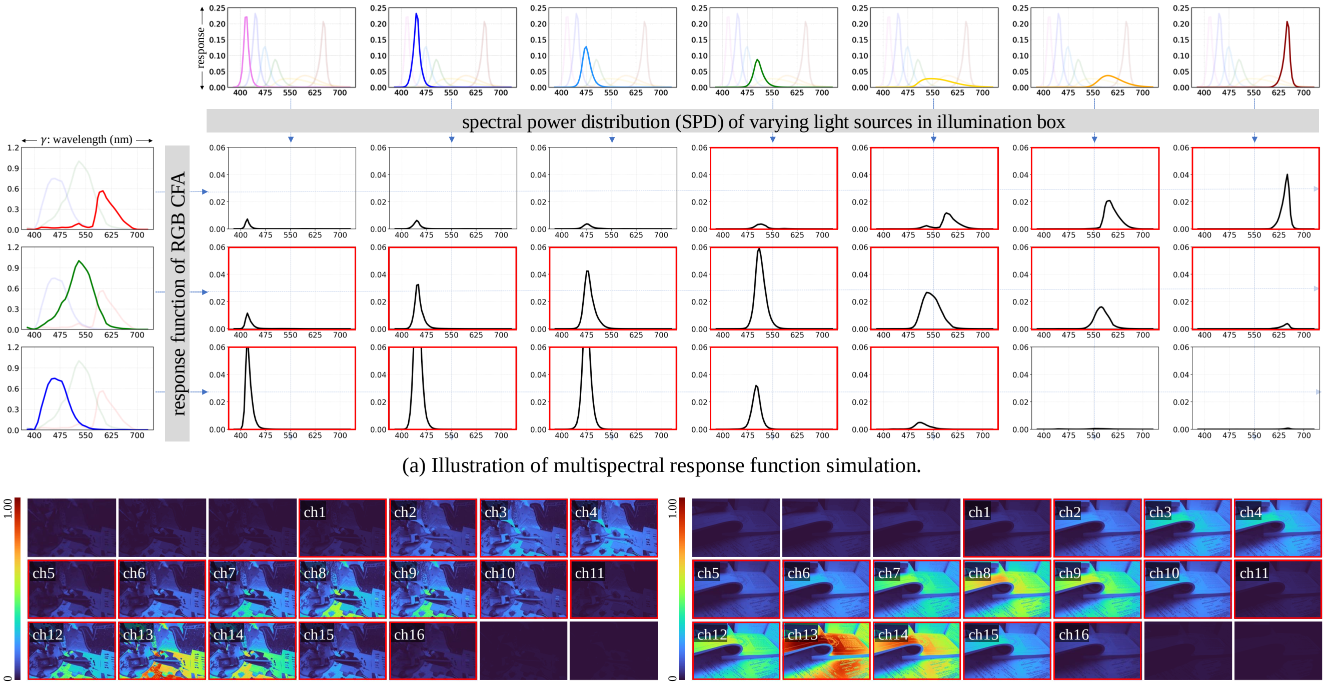


Figure 1. Overview of the data capturing pipeline: The system uses a Sony Alpha 1 camera with an RGB Bayer CFA sensor and pixel-shift mode to capture high-resolution ground-truth RGB and multispectral (MS) images via sub-pixel shifts. The camera, mounted on a linear stage actuator, captures staged scenes from different positions while maintaining a fixed baseline. RGB images are acquired in an illumination box simulating CIE D65 daylight, while MS images use a multispectral filter array (MSFA) under varied lighting. The dataset includes 502 quadruplets from 28 challenging scenes, featuring diverse staged setups.



(a) Illustration of multispectral response function simulation.

(b) Example MS channel images from Scene #05

(c) Example MS channel images from Scene #11

Figure 2. Illustration of 16-channel MS image acquisition in our imaging system. (a) shows the 4×4 MSFA simulation using and RGB CFA camera (first column) and varying light sources provided by a configurable illumination box (top row). Given the RGB response function $C^i_{rgb}(x, \gamma)$ of the camera and 7 distinct SPDs $L^j(\gamma)$ provided by configurable illumination box, we simulate 21 multispectral response functions as $C^k_{ms} = C^i_{rgb}(x, \gamma)L^j(\gamma)$ for each $i \in [1, 2, 3]$ and $j \in [1, 2, ..., 7]$. From these 21 response functions, 16 are selected to simulate a 4×4 MSFA (highlighted by red boxes). (b-c) are examples of MS images of different scenes captured using the multispactral response functions. The selected 16-channel images (highlighted by red boxes) are used to construct the ground-truth MS demosaic image $\hat{I}^{MS}$ in the proposed dual-camera RGB-MS dataset.

1.2. Noise Calibration

Given demosaiced RGB and MS images captured by our imaging system, we generate their corresponding mosaic images. To mitigate noise caused by the small pixel size of the sensor, we first downsample the pixelshift demosaiced images from 5640×8760 to 1440×2160. These downsampled clean images serve as our demosaiced ground-truth in our proposed dataset. Next, we apply synthetic noise and then mosaic the images using a 2×2 Bayer CFA for RGB and a 4×4 MSFA for MS to obtain the final mosaic images.

To simulate realistic sensor noise, we use a Poisson-Gaussian noise model [5, 10, 11]. Given the clean image $I \in \mathbb{R}^{H \times W \times N}$, noisy image $Y \in \mathbb{R}^{H \times W \times N}$ is modeled as:

$$Y_n(x) = I_n(x) + \epsilon_n(I_n(x)), \tag{2}$$

where n denotes the channel index and $\epsilon_n(I_n(x))$ represents the noise at pixel location x. The noise distribution is calibrated using heteroscedastic modeling, which accounts for the per-pixel signal dependency of photon noise. Mathematically:

$$\epsilon_n(I_n(x)) \sim \mathcal{N}(0, \sigma^2_n(I_n(x))), \text{where} \tag{3}$$

$$\sigma^2_n(I_n(x)) = \beta^1_n I_n(x) + \beta^2_n. \tag{4}$$

Here, $\sigma^2_n(I_n(x))$ represents the intensity-dependent noise variance. The parameter β^1_n models photon shot noise, proportional to the pixel intensity $I_n(x)$, while β^2_n accounts for intensity-independent electronic read noise.

Following the procedure outlined in [5, 10], we calibrate the noise parameters for each RGB CFA channel of the Sony Alpha 1 camera. To this end, we capture images of the X-Rite color chart under different exposures and ISO levels, and fit a linear model to the scatter plot of the calculated mean and variance pairs of pixel intensities at all homogeneous patches of the color chart images. This linear fit describes the heteroscedastic noise variance as a function of pixel intensity, determined separately for each color channel at different ISO values. In practice, our pipeline synthesizes noise using β^1_n and β^2_n calibrated at ISO 400. Nonetheless, it is worth mentioning that noise synthesis can be easily extended to other ISO levels, as mosaic images can be regenerated using the high-quality ground-truth RGB and MS demosaiced images in our dataset.

Once β^1_n and β^2_n are calibrated for the camera, we apply the noise to the clean RGB and MS demosaiced images captured by our system, based on their spectral channel and intensity value. Note that the noise model calibrated for

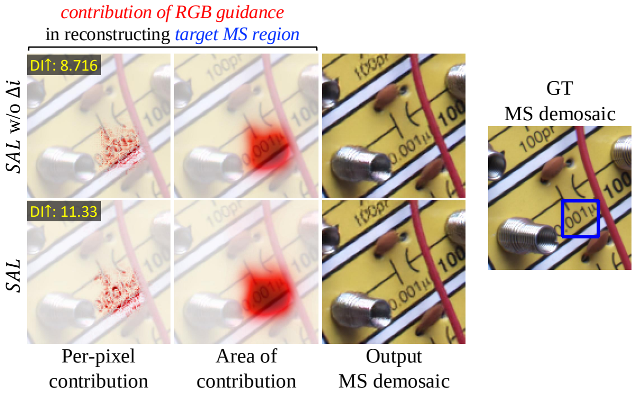


Figure 3. Qualitative analysis of the Spectral Alignment Layer (SAL). The figure compares results without spectral alignment offsets (SAL w/o Δi, top row) and with the offsets (SAL, bottom row). Columns represent the per-pixel contribution of RGB guidance, the area of contribution, and the reconstructed MS demosaiced output. The results indicate that SAL enhances the effectiveness of RGB guidance in reconstructing accurate MS images. The ground-truth MS demosaiced image is shown on the far right.

the RGB CFA is directly applicable to simulate MS images, which are simulated by combining RGB channel responses $C_{rgb}(x, \gamma)$ with varying SPDs $L^j(\gamma)$ from the illumination box (Eq. (1)). The noise variance depends only on the pixel intensity $I(x)$, which aggregates the contributions of SPDs through the image formation process (Eqs. (11) and (12) in the main paper). While SPDs can indirectly influence $I(x)$, the noise model itself is calibrated based on intensity and intrinsic sensor characteristics, making it agnostic to specific SPD. As such, the calibrated noise model remains valid for simulated MS images, ensuring consistency in noise synthesis regardless of variations in the illuminant or spectral composition, as long as the pixel intensities are preserved.

1.3. Color Conversion Matrix Calibration

To enable cross-spectral alignment in our MS demosaicing framework (Sec. 3.2 of the main paper), our dataset includes a pre-calibrated MS-to-RGB color conversion matrix. This matrix converts the MS image into RGB color space, ensuring spectral compatibility between the MS and RGB images during geometric alignment. The matrix is calibrated using RGB and MS images of the X-Rite Digital-SG color chart, which contains 140 color patches.

The calibration process extracts average patch intensities, resulting in two matrices: $\mathbf{A} \in \mathbb{R}^{140\times16}$, representing the multispectral values, and $\mathbf{B} \in \mathbb{R}^{140\times3}$, representing the RGB values. The color conversion matrix $C \in \mathbb{R}^{16\times3}$, which maps 16-channel MS to 3-channel RGB values, is computed using least-squares optimization:

$$C = \underset{C}{\text{argmin}} \|\mathbf{A} \times C - \mathbf{B}\|^2. \tag{5}$$

The resulting conversion matrix C is crucial in cross-spectral disparity estimation (Eq. (3) in the main paper), where it transforms the MS image into the RGB color space before disparity estimation, enabling geometric alignment between MS and RGB images in the proposed framework.

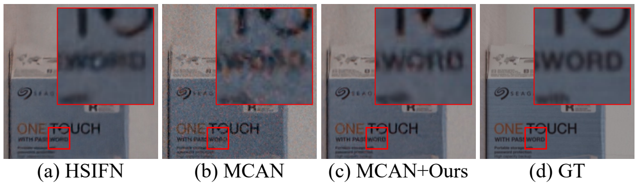

(a) HSIFN (b) MCAN (c) MCAN+Ours (d) GT

Figure 4. Qualitative comparison of RGB-guided MS demosaicing on the HS dataset [7]. Trained solely on our dataset, MCAN + Ours recovers sharper spatial details and cleaner spectral content than MCAN and HSIFN, demonstrating the generalization capability of our method across varying MS imaging conditions.

2. Analysis on MS Demosaicing Framework

2.1. Generalization

While our method is primarily trained and evaluated on our large scale dataset—due to the availability of paired RGB-MS mosaics with high-fidelity GTs—we also conducted cross-dataset experiments on the HS dataset [7], which contains 60 HS-HS image pairs. Although it lacks raw mosaics, we simulate MS and RGB mosaics by projecting the demosaiced HS images onto our MSFA and CFA spectral response functions. This enables compatibility but introduces a domain gap due to differences in sensor characteristics. Simulated MS/RGB images can deviate from real sensor responses due to limitations in HS data precision, spectral response calibration, sensor nonlinearities, and noise characteristics [1]. Despite these challenges, our model generalizes well: MCAN + Ours achieves 42.48dB PSNR (24.03M, 2.66T MACs), outperforming MCAN (39.40dB, 5.24M, 0.91T) and the high-capacity HSIFN (40.27dB, 90.21M, 18.79T), demonstrating the effectiveness of our fusion strategy. Fig. 4 illustrates the qualitative advantages of our method.

2.2. Effect of Spectral Alignment Layer (SAL)

We analyze the effectiveness of the Spectral Alignment Layer (SAL) in enhancing MS image demosaicing by leveraging RGB guidance. Figure 3 illustrates the contribution of RGB guidance to reconstructing a target MS region, comparing results with and without SAL. The top row shows results without the spectral alignment offsets Δi (Eq. (7) in the main paper), while the bottom row includes the proposed SAL (the last row in the table). The two models provides multi-scale RGB features $f_l'^{RGB}$ to the fusion network $\mathcal{F}$ (Eq. (5) in the main paper). The per-pixel contribution and the area of contribution indicate that SAL enables more effective integration of RGB guidance, leading to improved MS demosaicing quality, as reflected in the sharper and more accurate output.

Table 1 presents quantitative results in terms of the Diffusion Index (DI) [6], which measures the range of contributed RGB pixels during MS demosaicing. Higher DI scores indicate better utilization of RGB guidance. Results show that incorporating SAL consistently improves DI across all MS channels, with an average gain of 2.138, demonstrating the importance of spectral alignment in achieving higher fidelity MS reconstructions.

2.3. Optical Flow Visualization

The scenes in our dataset have objects at various depths and thus we utilize optical flow to perform alignment within our cross-spectral disparity estimation module. In Figure 5, we visualize the optical flow field that warps the RGB features into alignment with the intermediate MS features. We visualize the intermediate MS image and demosaiced RGB image and the optical flow computed. The warped RGB visualization is a simplification of our SAL, which warps features across multiple scales from the RGB image. The optical flow between all images should be in the same direction since the cameras are separated by a fixed distance. However, the magnitude of the shift varies with scene depth–closer objects exhibit larger flow values, while farther objects have smaller flow values.

2.4. RGB to MS Reconstruction

In our experiments (Sec. 5 of the main paper), we demonstrate the effectiveness of using the RGB mosaic as guidance for MS restoration. However, a natural question arises: can we directly reconstruct the MS image from the RGB mosaic alone, bypassing the need for the MS mosaic? To verify our approach, we assess our model against the earlier RGB-to-HS image reconstruction model, MSTPP$_{1\times}^{\text{rgb2ms}}$ [2], which is adapted to process RGB mosaic images and produce demosaiced MS images for the 1× MS demosaicing task (*i.e*., Scenario 1 in Sec. 5.2 of the main paper).

For our method, we prepare three model variants. The first variant, NAFNet$_{1\times}^{\text{rgb2ms}}$, uses NAFNet [3] as the backbone MS restoration network $\mathcal{D}_{MS}$. It takes only the RGB mosaic image $I_{2\times 2}^{RGB}$ as input and directly produces the MS demosaiced image I^{MS} as output. The second variant, NAFNet$_{1\times}$, also uses NAFNet as the backbone but instead takes the MS mosaic $I_{4\times 4}^{MS}$ as input to reconstruct I^{MS}. The third variant, NAFNet$_{1\times}$ + Ours, integrates our proposed modules to leverage RGB guidance for MS demosaicing. These modules include the RGB demosaicing network $\mathcal{D}_{RGB}$ (Eq. (2) in the main paper) and the cross-spectral fusion module (Sec. 3.2 of the main paper).

For training, MSTPP$_{1\times}^{\text{rgb2ms}}$ and our first model variant, NAFNet$_{1\times}^{\text{rgb2ms}}$, are trained using RGB mosaic images $I_{2\times 2}^{RGB}$, whereas our second model variant, NAFNet$_{1\times}$, is trained to handle MS mosaic images $I_{4\times 4}^{MS}$. The training process employs L2 loss between predicted MS demosaic images with the ground-truth $\hat{I}^{MS}$ (Eq. (8) of the main paper). Note that the $I_{2\times 2}^{RGB}$ images are geometrically aligned with ground-truth MS images, as they are captured under CIE D65 daylight illumination from the same camera position as $\hat{I}^{MS}$. The third variant, NAFNet$_{1\times}$ + Ours, follows the full training pipeline outlined in Section 3.3 of the main paper, which incorporates RGB guidance through fusion to improve MS restoration.

Target MS channel	DI [6] w.r.t. target MS channel (↑) SAL w/o Δi	 SAL	Gain
1	9.035	**11.22**	2.185
2	8.500	**9.976**	1.476
3	8.962	**8.964**	0.002
4	8.606	**9.753**	1.146
5	8.757	**10.63**	1.874
6	7.388	**9.518**	2.130
7	7.296	**9.103**	1.807
8	10.04	**10.12**	0.088
9	8.203	**9.861**	1.658
10	8.728	**9.989**	1.260
11	9.215	**12.31**	3.095
12	8.049	**9.033**	0.984
13	7.823	**8.094**	0.270
14	7.208	**8.310**	1.102
15	8.012	**9.837**	1.825
16	8.487	**10.77**	2.283
DI w.r.t all MS channels	7.373	**9.511**	2.138

Table 1. Quantitative results for 1× MS demosaicing, showing the effect of SAL in terms of the Diffusion Index (DI) [6]. The DI measures the range of involved RGB pixels during MS restoration. The reported values represent the average Diffusion Index (DI), calculated by selecting two random target regions from each of the 103 test images in the proposed RGB-MS dataset, resulting in a total of 206 target regions.

Table 2 summarizes the results. The table highlights the clear advantage of using the MS mosaic input for MS reconstruction (first and second vs. third rows of the table). Moreover, providing RGB guidance with our proposed modules achieves the best MS restoration performance (fourth row), as the high-fidelity details from the RGB mosaic are effectively fused during MS restoration.

We also compare MSTPP$_{1\times}^{\text{rgb2ms}}$ and NAFNet$_{1\times}^{\text{rgb2ms}}$ with NAFSR$_{4\times}$ [4], a variant of NAFNet designed for super-resolution and trained for the 4× MS demosaicing task, which reconstructs MS images from 4× downsampled MS mosaics (Scenario 2 in Sec. 5.2 of the main paper). As expected, MSTPP$_{1\times}^{\text{rgb2ms}}$ and NAFNet$_{1\times}^{\text{rgb2ms}}$ outperform NAFSR$_{4\times}$ across all metrics, as the latter relies on lower-resolution inputs. However, when NAFSR$_{4\times}$ is combined with our proposed modules (last row of Table 2), it achieves competitive PSNR and SSIM scores compared to NAFNet$_{1\times}^{\text{rgb2ms}}$, while significantly improving the SAM score, which quantifies spectral fidelity. This demonstrates

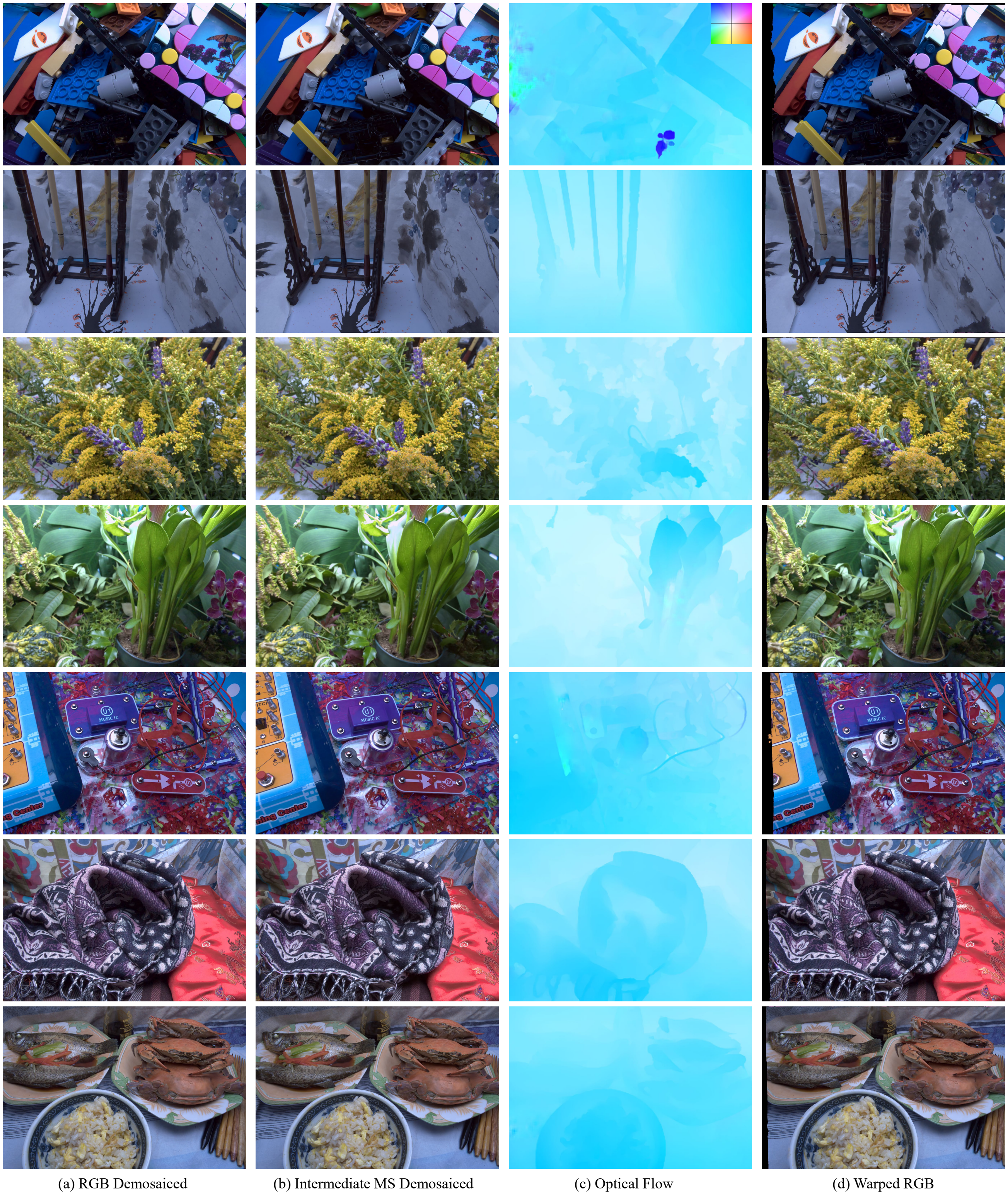

(a) RGB Demosaiced (b) Intermediate MS Demosaiced (c) Optical Flow (d) Warped RGB

Figure 5. Visualization of optical flow and warping on various scenes in our test dataset. In the first two columns, we visualize the demosaiced RGB and the intermediate MS image (converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point), that are used to compute the flow. In the last two columns, we visualize the optical flow and the RGB image backwards warped into alignment with the MS image. Note that we visualize the warped RGB image, but in practice our SAL warps *features* across multiple scales.

Model	PSNR↑	SSIM↑	SAM↓	Params (MB)	MACs (T)
$\text{MSTPP}^{\text{rgb2ms}}_{1\times}$ [2]	37.96	0.9746	4.430	85.81	12.36
$\text{NAFNet}^{\text{rgb2ms}}_{1\times}$	37.94	0.9734	4.370	111.25	0.78
$\text{NAFNet}_{1\times}$ [3]	40.89	0.9766	2.604	111.25	0.78
$\text{NAFNet}_{1\times}$ + Ours	**41.92**	**0.9811**	**2.422**	130.03	2.53
$\text{NAFSR}_{4\times}$ [4]	32.98	0.9173	4.736	59.19	2.66
$\text{NAFSR}_{4\times}$ + Ours	37.67	0.9641	3.576	77.98	4.41

Table 2. Quantitative comparison for MS demosaicing.

the efficacy of the proposed RGB-guided MS restoration scheme, even in challenging super-resolution settings.

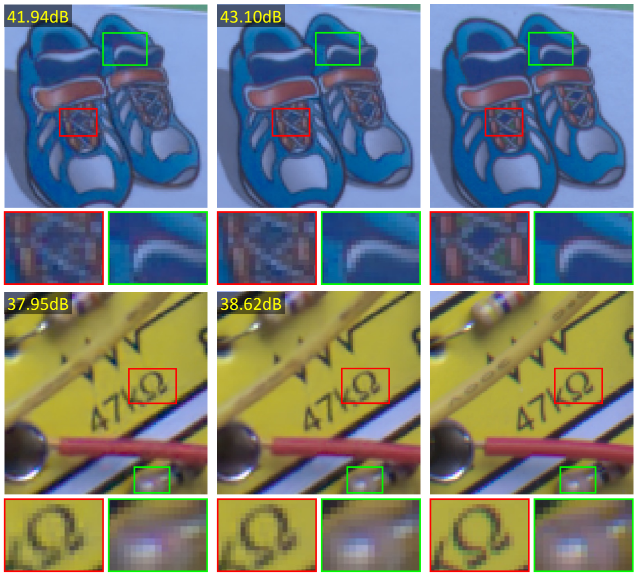
2.5. RGB Demosaicing using MS Reference

While the primary focus of this work is on leveraging RGB guidance for MS restoration tasks in dual-camera setups with RGB and MS sensors, the complementary nature of the MS sensor motivates exploring the inverse scenario: leveraging MS guidance to enhance RGB demosaicing. Despite the lower spatial fidelity of MS mosaics due to their inherent low-resolution nature, they can potentially capture spectral details that cannot be captured by the RGB CFA sensor [12]. Furthermore, MS sensor provide richer spectral diversity, which can be utilized during RGB demosaicing tasks for reconstructing accurate colors [9, 14]. Our proposed framework and dual-camera RGB-MS dataset are well-suited for extending to this task, demonstrating their flexibility in addressing different restoration scenarios.

To validate this idea, we adapt our proposed MS demosaicing framework by prioritizing RGB restoration in the fusion stage (Sec. 3.2 of the main paper). Specifically, the demosaiced RGB image I'^{RGB} is used as the primary input to the fusion network $\mathcal{F}$, while the intermediate multi-scale MS feature map $f_l'^{MS}$ is refined by the spectral alignment layer (SAL) and provided as auxiliary MS guidance to $\mathcal{F}$. This adjustment allows the fusion network to generate enhanced RGB demosaiced image I^{RGB} by leveraging the spectral diversity of the MS features.

We evaluate the effectiveness of MS-guided RGB demosaicing by comparing three model variants. The baseline model, NAFNet [3], processes only RGB mosaic images $I^{RGB}_{2\times2}$ without MS guidance. To ensure a fair comparison, we also evaluate a capacity-increased version, NAFNet-L. Finally, the proposed method, NAFNet+Ours, incorporates MS guidance during the fusion stage, integrating $f_l'^{MS}$ to enhance RGB reconstruction. All models are trained using the RGB demosaicing loss (Eq. (9) in the main paper) on paired RGB mosaic images $I^{RGB}_{2\times2}$ and their corresponding ground-truth RGB demosaiced images $\hat{I}^{RGB}$.

Table 3 presents the quantitative results. Compared to NAFNet, NAFNet integrated with our modules achieves consistent improvement across all metrics (first vs. third rows of the table), demonstrating the benefit of incorporating MS guidance in RGB restoration. Compared to NAFNet-L, which benefits from increased model capacity, our method shows better performance, illustrating effectiveness of our method in utilizing MS guidance for enhancing RGB reconstruction quality. Figure 6 further illustrates the benefits of the proposed method. The zoomed-in regions show how MS guidance contributes to improved detail recovery and color accuracy. Specifically, compared to NAFNet-L, our method recovers finer spatial details (red boxes in first vs. second columns) and addresses color inaccuracies caused by limited spectral diversity of RGB CFA sensor (green boxes), validating the effectiveness of MS-guided RGB demosaicing. Furthermore, the results highlight the flexibility of our dual-camera RGB-MS dataset and framework, indicating their potential to support both MS restoration and RGB reconstruction tasks.



(a) NAFNet-L (b) NAFNet + Ours (c) GT

Figure 6. Qualitative comparison of RGB demosaicing: NAFNet-L processes only the RGB mosaic image as input, whereas NAFNet+Ours incorporates MS features as guidance during demosaicing. The zoomed-in cropped patches in the red and green boxes demonstrate the advantages of using MS guidance: enhanced detail (red box) and improved color accuracy (green box).

3. More Details and Discussion on Experiments

In this section, we provide additional details comparing our model with the previous approaches (Sec. 5.2 of the main paper).

3.1. Architectural Modification

Adapting Alignment For both DCT [8] and HSIFN [7], we perform alignment to ensure a fair comparison. DCT requires alignment, so we use our pre-alignment module to provide an aligned RGB image alongside the MS mosaic. HSIFN includes its own alignment network, which requires

Model	PSNR↑	SSIM↑	SAM↓	Params (MB)	MACs[1] (T)
NAFNet [3]	44.97	0.9844	1.566	111.23	0.77
NAFNet-L	45.06	0.9847	1.546	158.57	1.49
NAFNet + Ours	**45.82**	**0.9874**	**1.450**	130.03	2.51

Table 3. Quantitative comparison for RGB demosaicing

two RGB images; we use this network but supply RGB inputs from our pre-alignment module.

Upsampling Module For Scenario 2 of the Sec. 5.2 of the main paper, which aims for the 4× MS demosaicing task, the baseline MS demosaicing networks $\mathcal{D}_{MS}$ are modified to reconstruct high-resolution MS images from low-resolution MS mosaics. For NAFSR [3], we adapt NAFSSR [4], originally designed for stereo super-resolution, by removing its stereo-specific cross-attention modules to accommodate the single mosaic input in our task. For MCAN and Restormer, we replace the final convolution layer, which produces a demosaiced image, with a feature extraction layer, followed by an upsampling module that generates high-resolution MS demosaiced images. The upsampling module comprises convolutional and pixel-shuffle [13] layers, with its architecture detailed in Table 4.

Discussion on HSIFN While both our method and HSIFN [7] use color mapping and optical flow for alignment, they differ in fusion strategy and backbone design. HSIFN applies spatial attention between warped RGB and RGB-mapped HS features prior to decoding, whereas our method adopts channel attention within the decoder [3], which is well suited for restoration. For alignment, we use deformable convolutions, which outperform the direct warping approach used in HSIFN (Table 1). Additionally, our model is more efficient in terms of computational cost (2.53T vs. 18.79T; Table 2).

Type	Input	Act	K	Ch	S	Output
conv	$f'^{MS}_{l=1}(I^{MS}_{4\times4})$	LReLU	3×3	$c \times 4$	1	Conv$_1$
PixShfl	Conv$_1$	-	-	c	-	Up$_1$
conv	Up$_1$	LReLU	3×3	$c \times 4$	1	Conv$_2$
PixShfl	Conv$_2$	-	-	c	-	Up$_2$
conv	Up$_2$	LReLU	3×3	c	1	Final

Table 4. Architecture of the upsampling module for 4× MS demosaicing. Abbreviations: Act = Activation, K = Kernel size, Ch = Channels, S = Stride, PixShfl = PixelShuffle, Up = Upsampled. The input $f'^{MS}_{l=1}(I^{MS}_{4\times4})$ represents the final feature map from the MS demosaicing network $\mathcal{D}_{MS}$ at scale level $l = 1$.

4. Additional Qualitative Results

We present qualitative results on the proposed dual-camera RGB-MS test set for the 1× MS demosaicing task (Scenario 1 in Sec. 5.2 of the main paper) in Figs. 7 to 16, and for the 4× MS demosaicing task (Scenario 2 in Sec. 5.2 of the main paper) in Figs. 17 to 26. The visualized results include MS demosaics converted to the sRGB color space, MS demosaic averaged across the channel dimension, per-channel MS demosaics, and error maps computed between the restored and ground-truth MS demosaiced images.

References

[1] 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. 4

[2] 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. 5, 7

[3] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 5, 7, 8

[4] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: Stereo image super-resolution using nafnet. In *CVPRW*, 2022. 5, 7, 8

[5] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008. 3

[6] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *CVPR*, 2021. 5

[7] 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. 4, 7, 8

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

[9] Muyao Niu, Zhihang Zhong, and Yinqiang Zheng. NIR-assisted video enhancement via unpaired 24-hour data. In *ICCV*, 2023. 7

[10] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *CVPR*, 2017. 3

[11] Guocheng Qian, Yuanhao Wang, Jinjin Gu, Chao Dong, Wolfgang Heidrich, Bernard Ghanem, and Jimmy S. J. Ren. Rethinking learning-based demosaicing, denoising, and super-resolution pipeline. In *ICCP*, 2019. 3

[12] 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. 7

[13] 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

[14] 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. 7

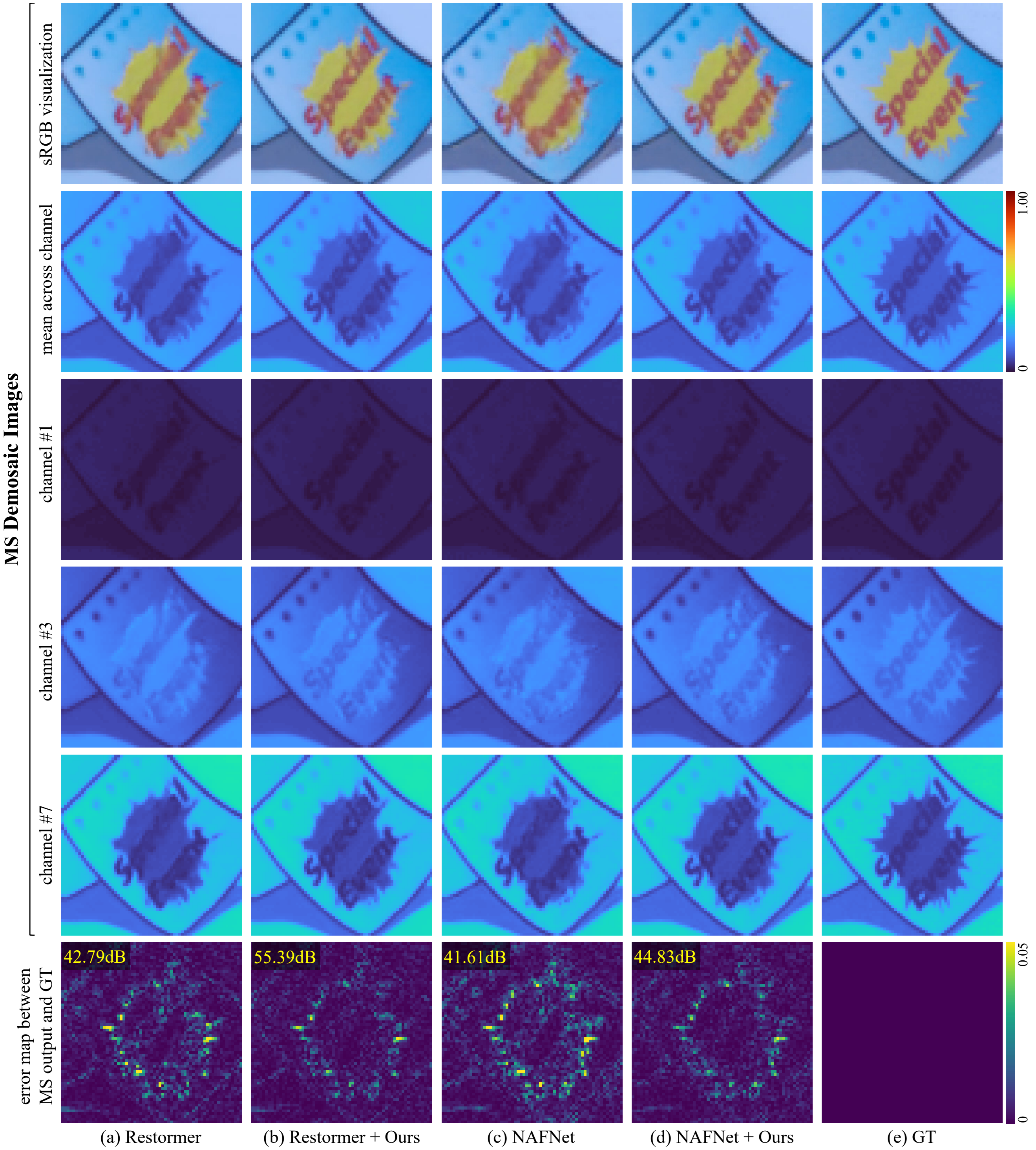


Figure 7. **Qualitative comparison of 1× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 1st, 3rd, and 7th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

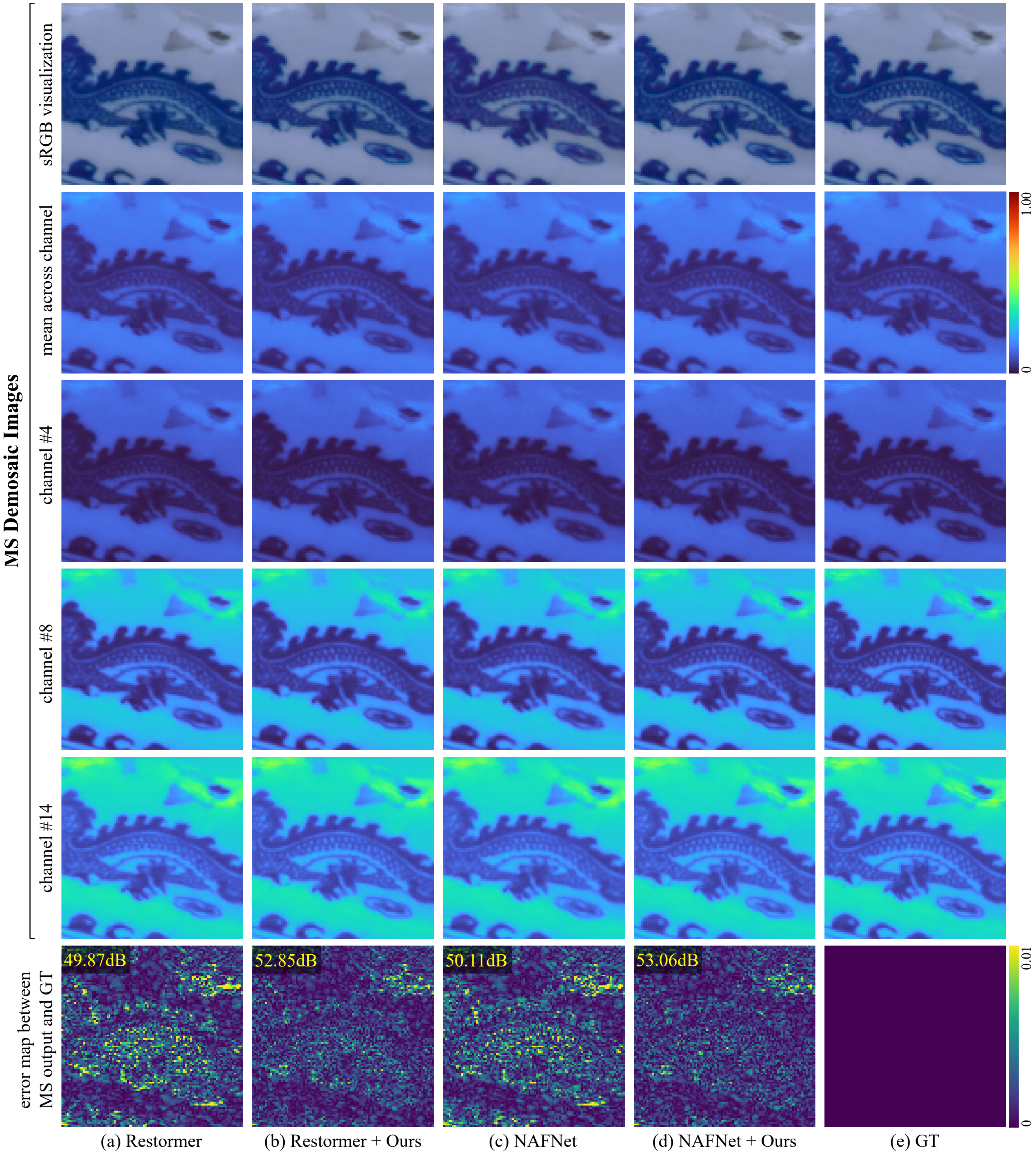


Figure 8. **Qualitative comparison of 1× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 4th, 8th, and 14th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

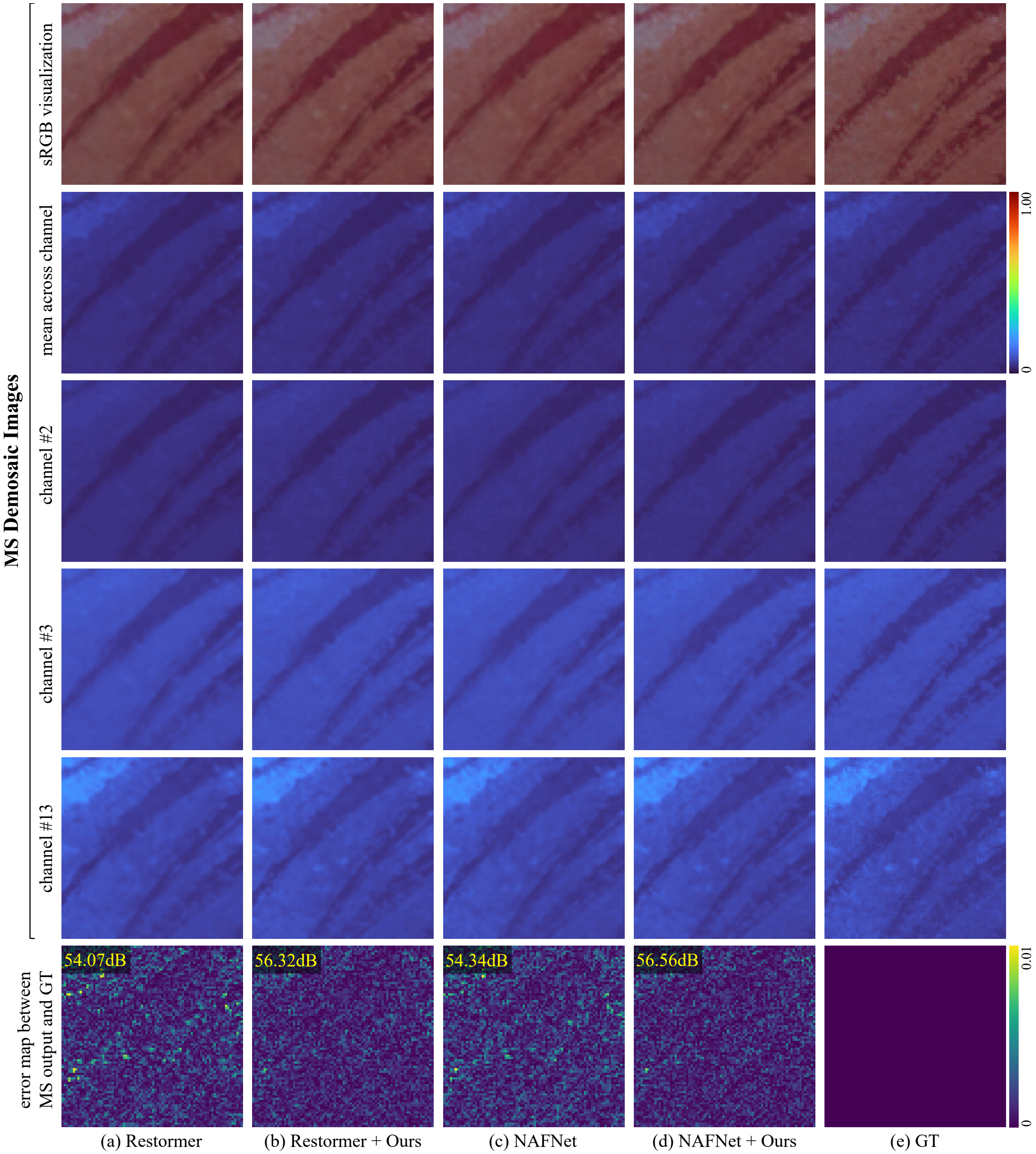


Figure 9. **Qualitative comparison of 1× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 2nd, 3rd, and 13th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

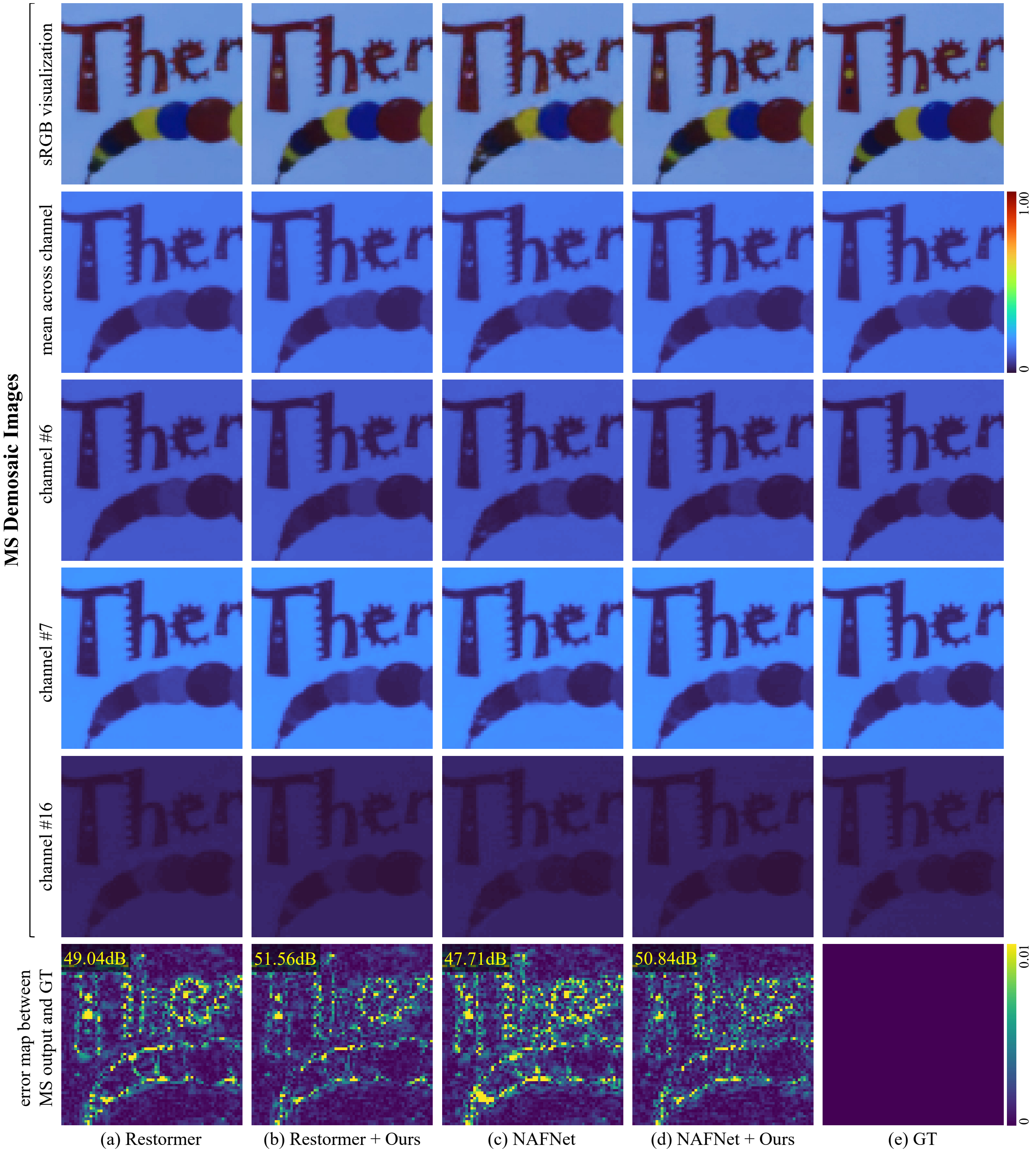


Figure 10. **Qualitative comparison of 1× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 6th, 7th, and 16th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

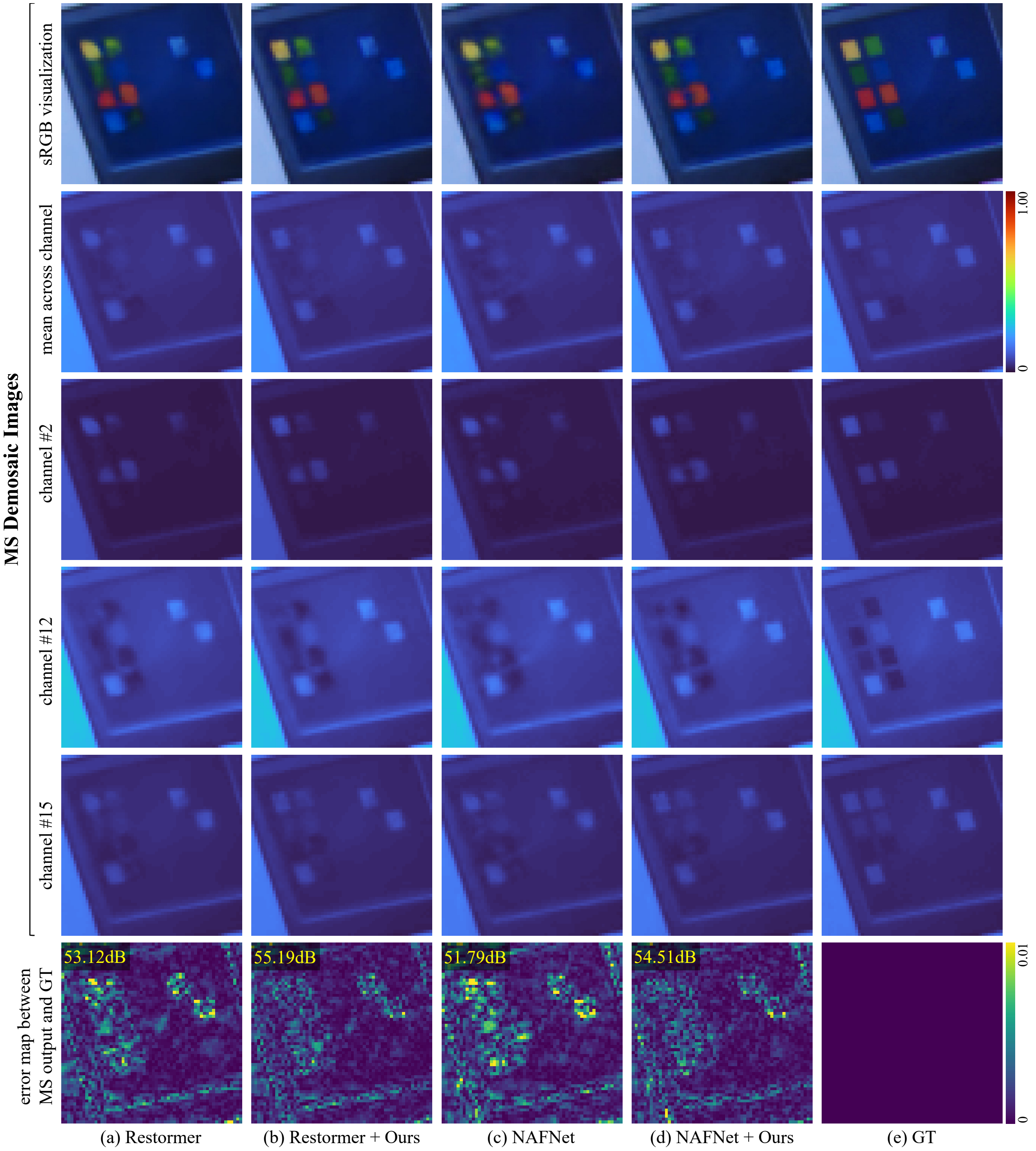


Figure 11. **Qualitative comparison of 1× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 2nd, 12th, and 15th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

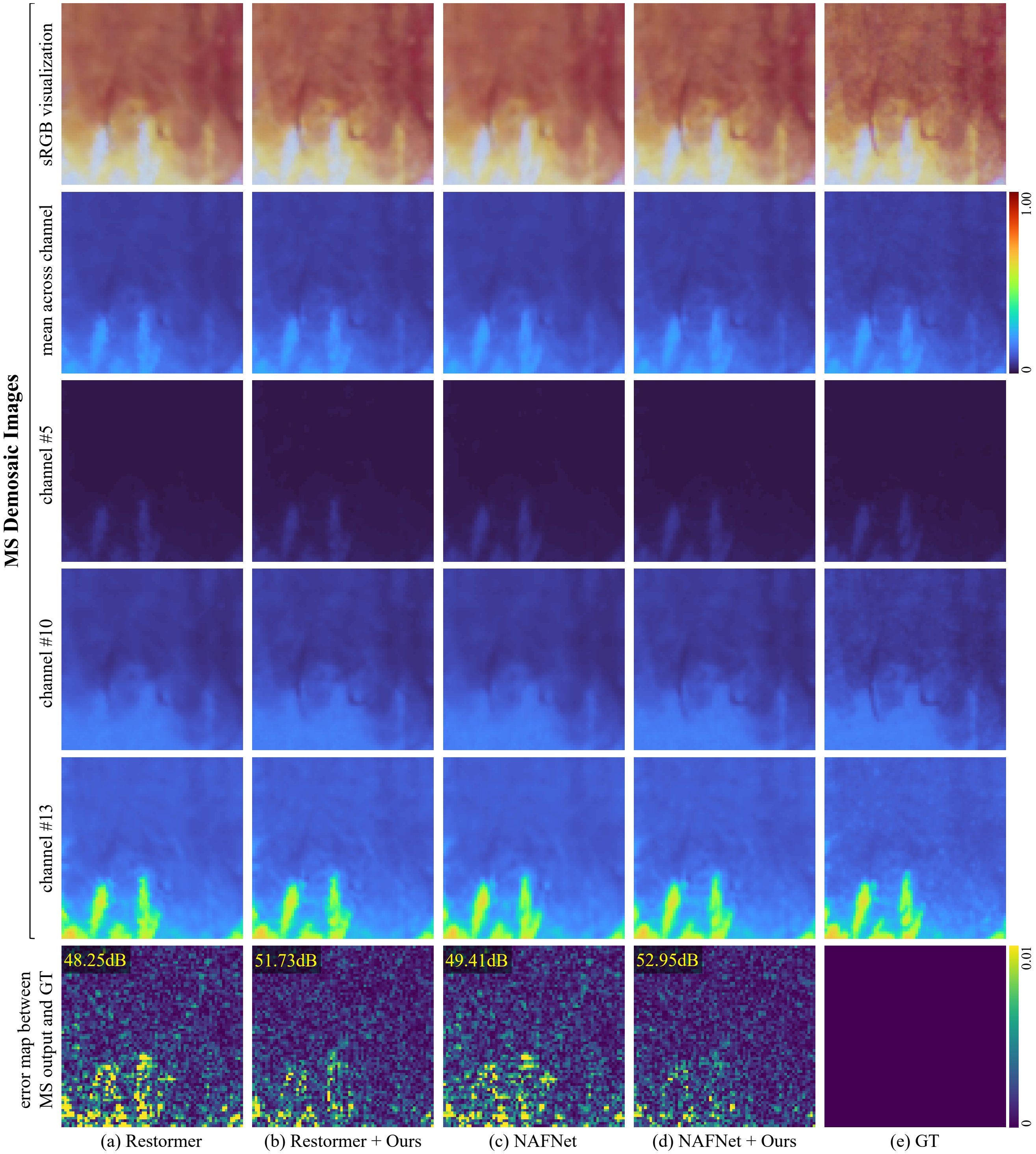


Figure 12. **Qualitative comparison of 1× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 5th, 10th, and 13th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

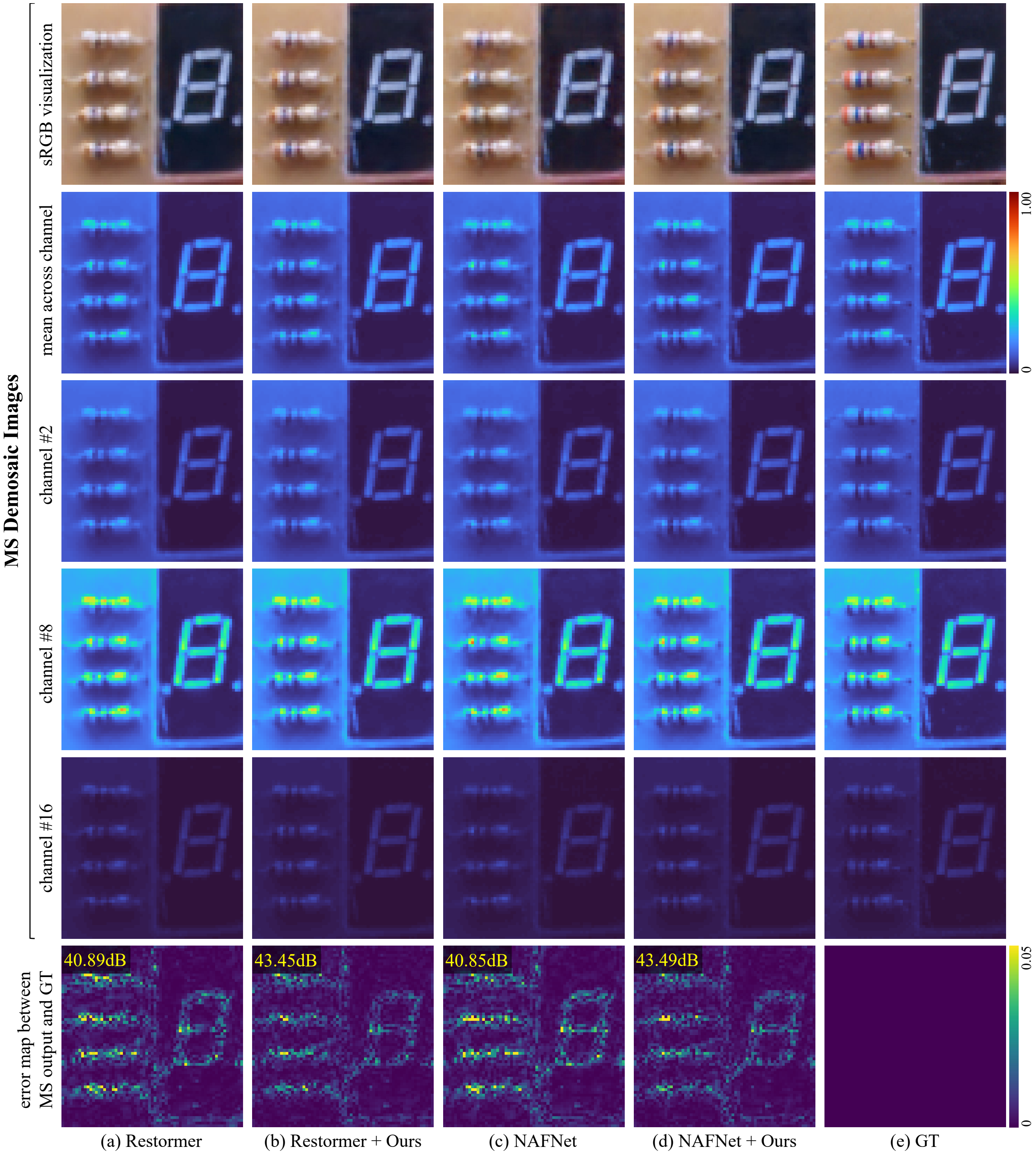


Figure 13. **Qualitative comparison of 1× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 2nd, 8th, and 16th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

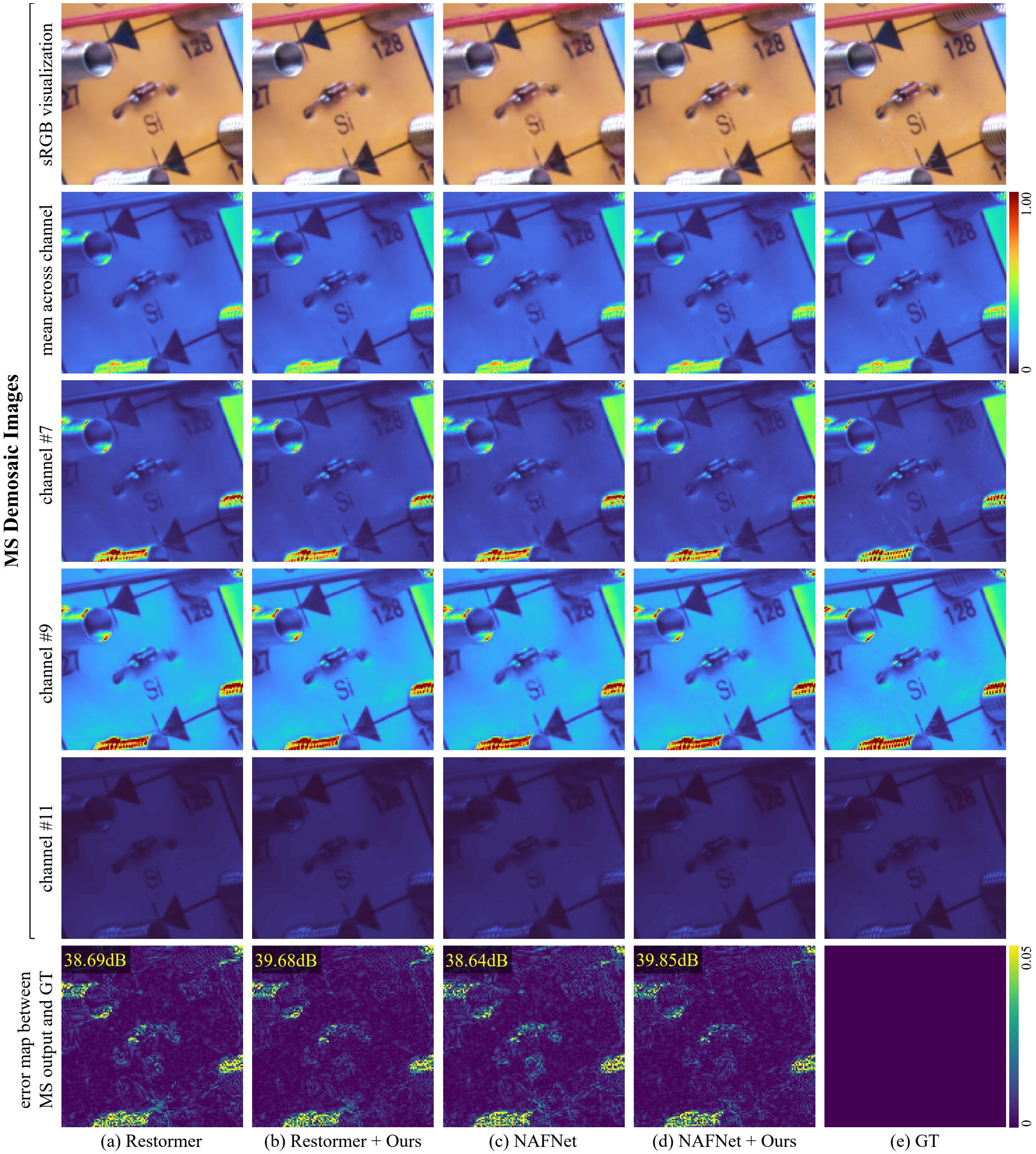


Figure 14. **Qualitative comparison of 1× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 7th, 9th, and 11th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

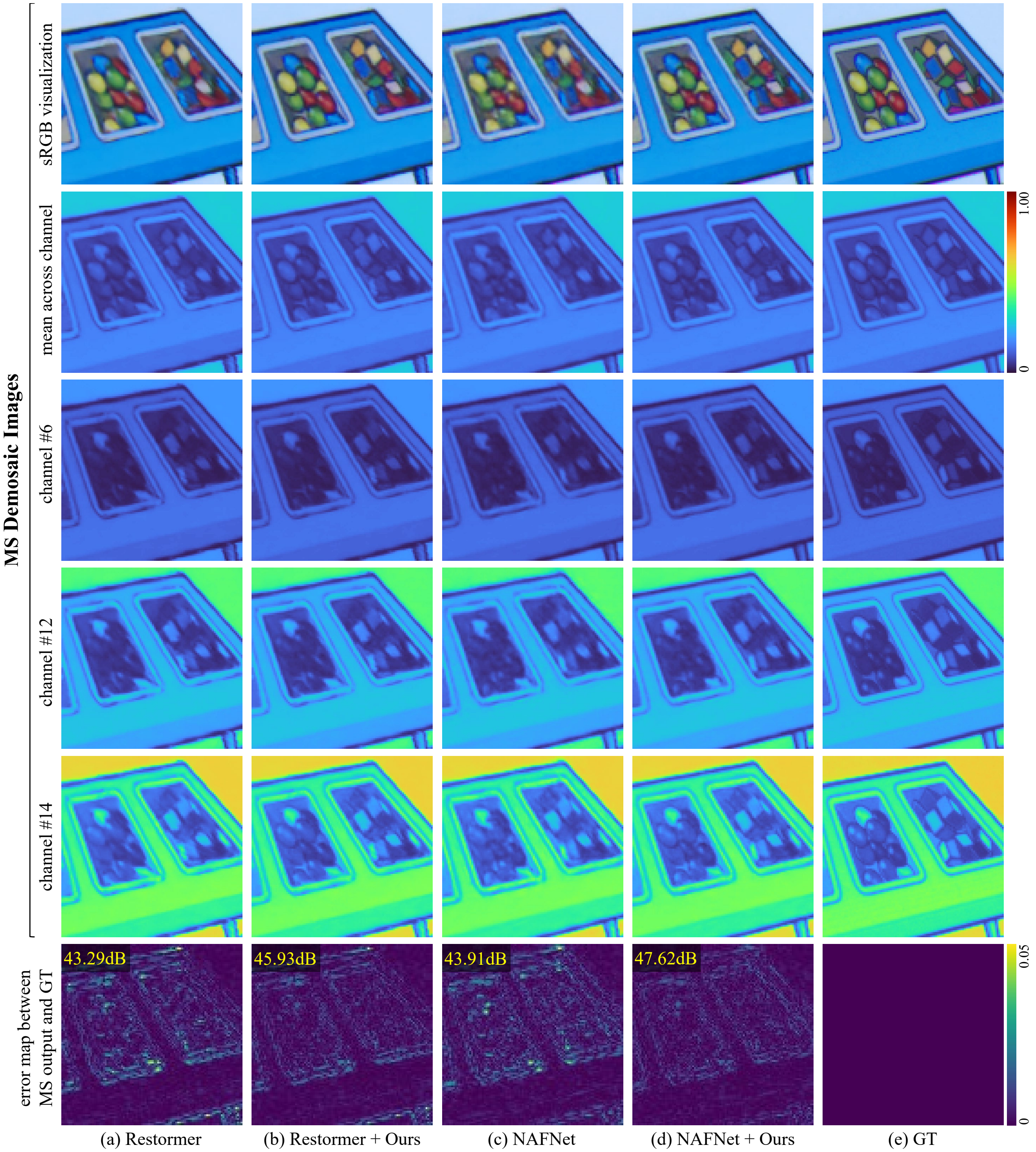


Figure 15. **Qualitative comparison of 1× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 6th, 12th, and 14th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

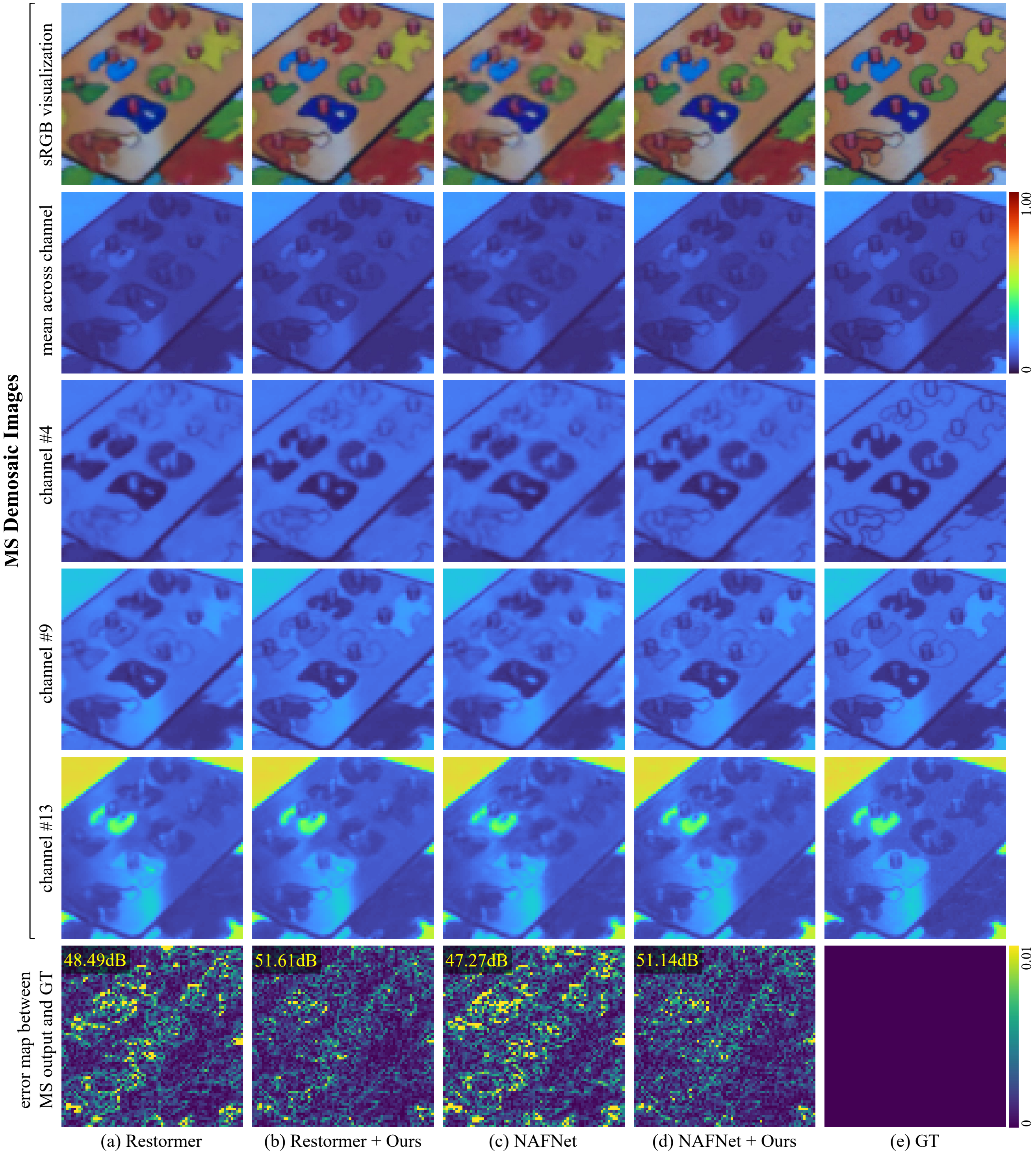


Figure 16. **Qualitative comparison of 1× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 4th, 9th, and 13th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

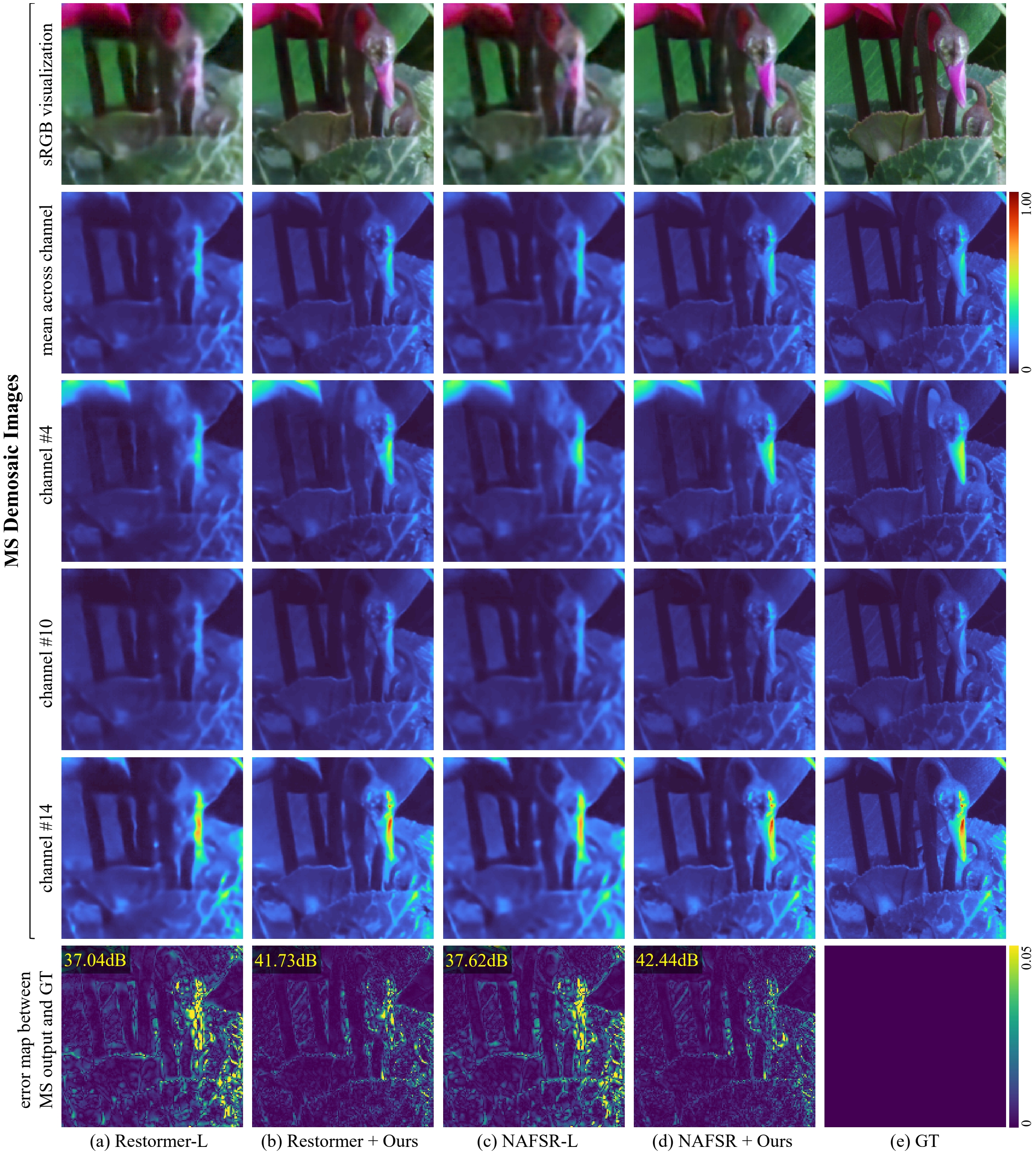


Figure 17. **Qualitative comparison of 4× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 4th, 10th, and 14th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

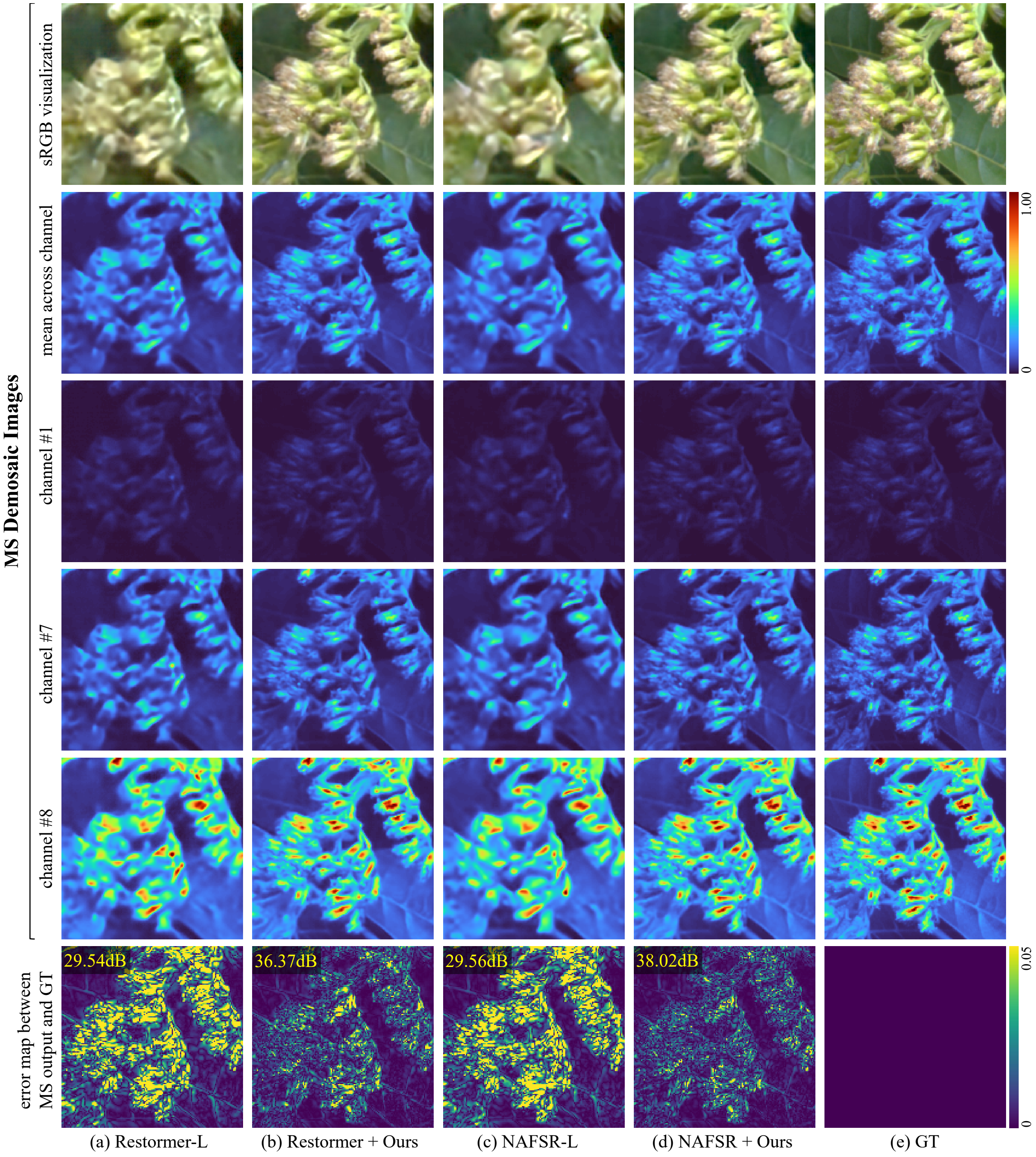


Figure 18. **Qualitative comparison of 4× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 1st, 7th, and 8th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

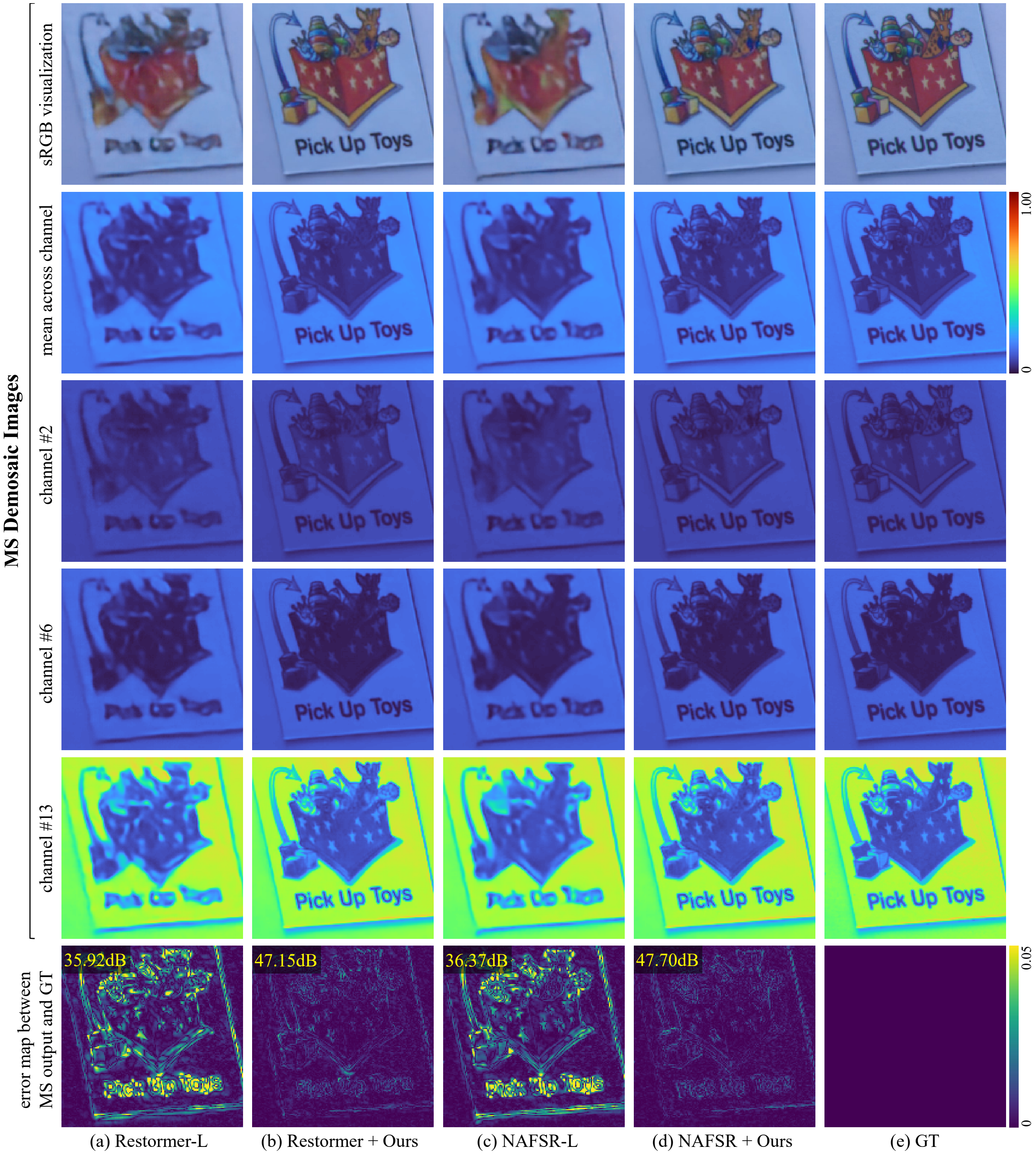


Figure 19. **Qualitative comparison of 4× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 2nd, 6th, and 13th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

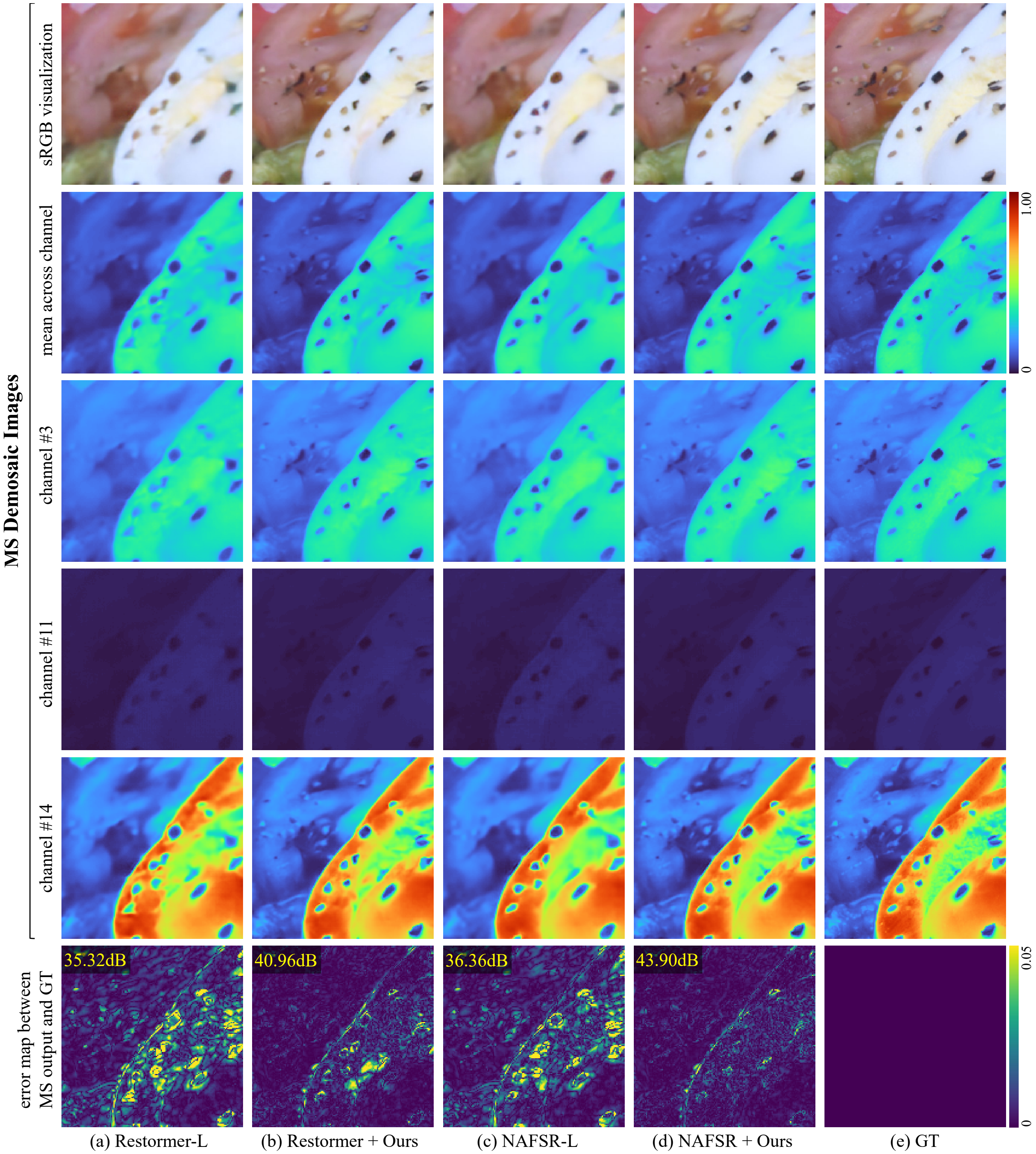


Figure 20. **Qualitative comparison of 4× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 3rd, 11th, and 14th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

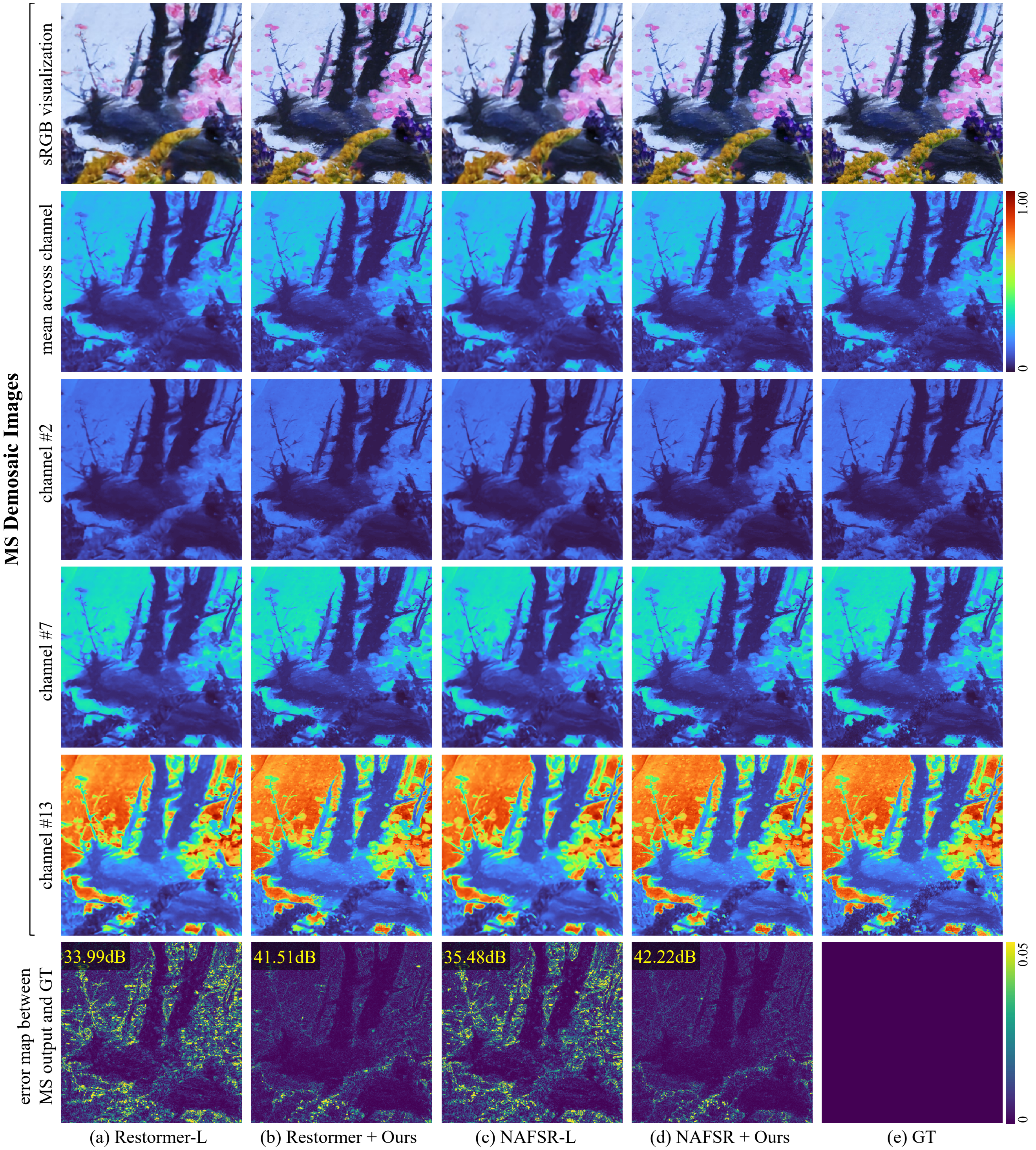


Figure 21. **Qualitative comparison of 4× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 2nd, 7th, and 13th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

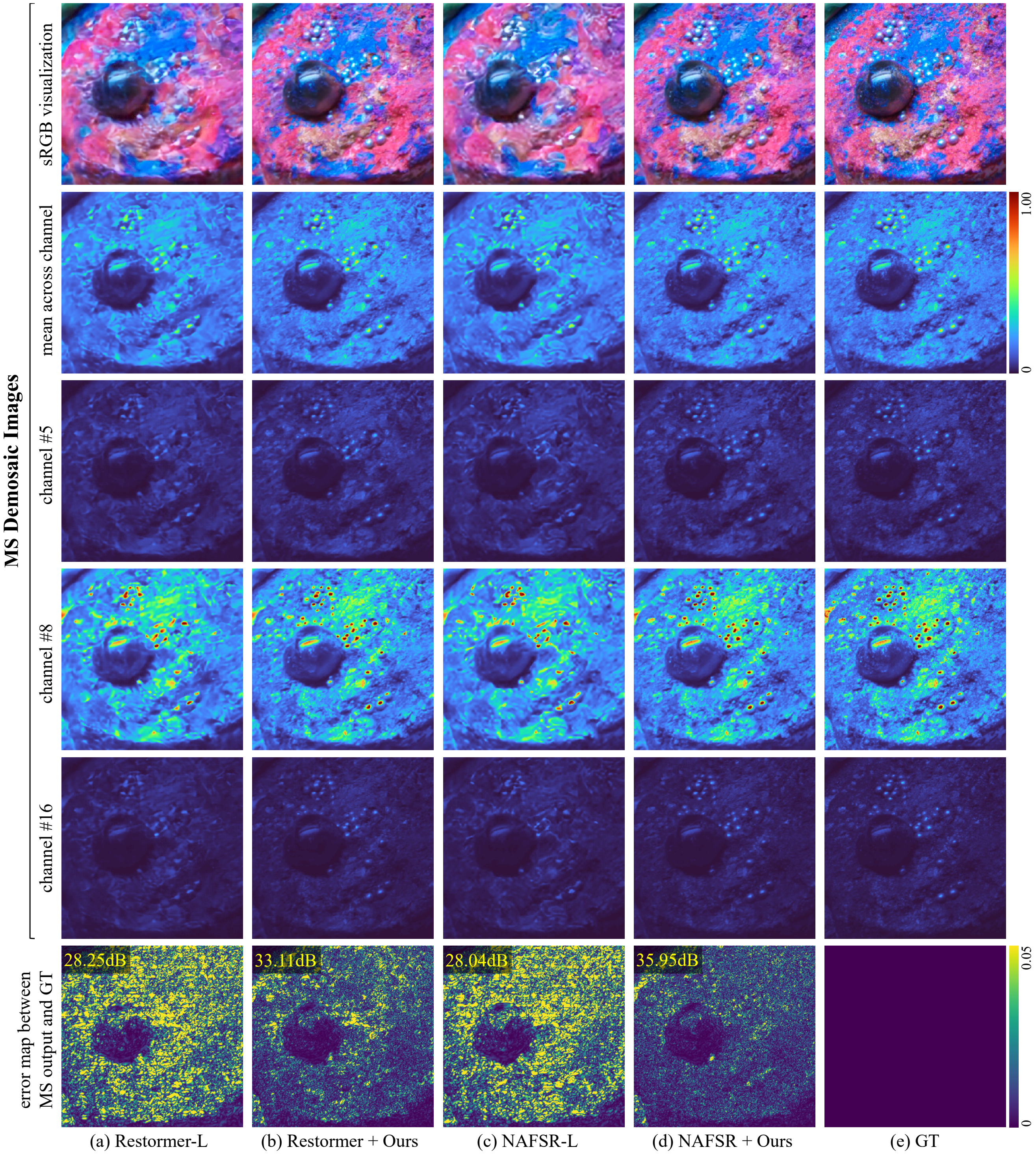


Figure 22. **Qualitative comparison of 4× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 5th, 8th, and 16th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

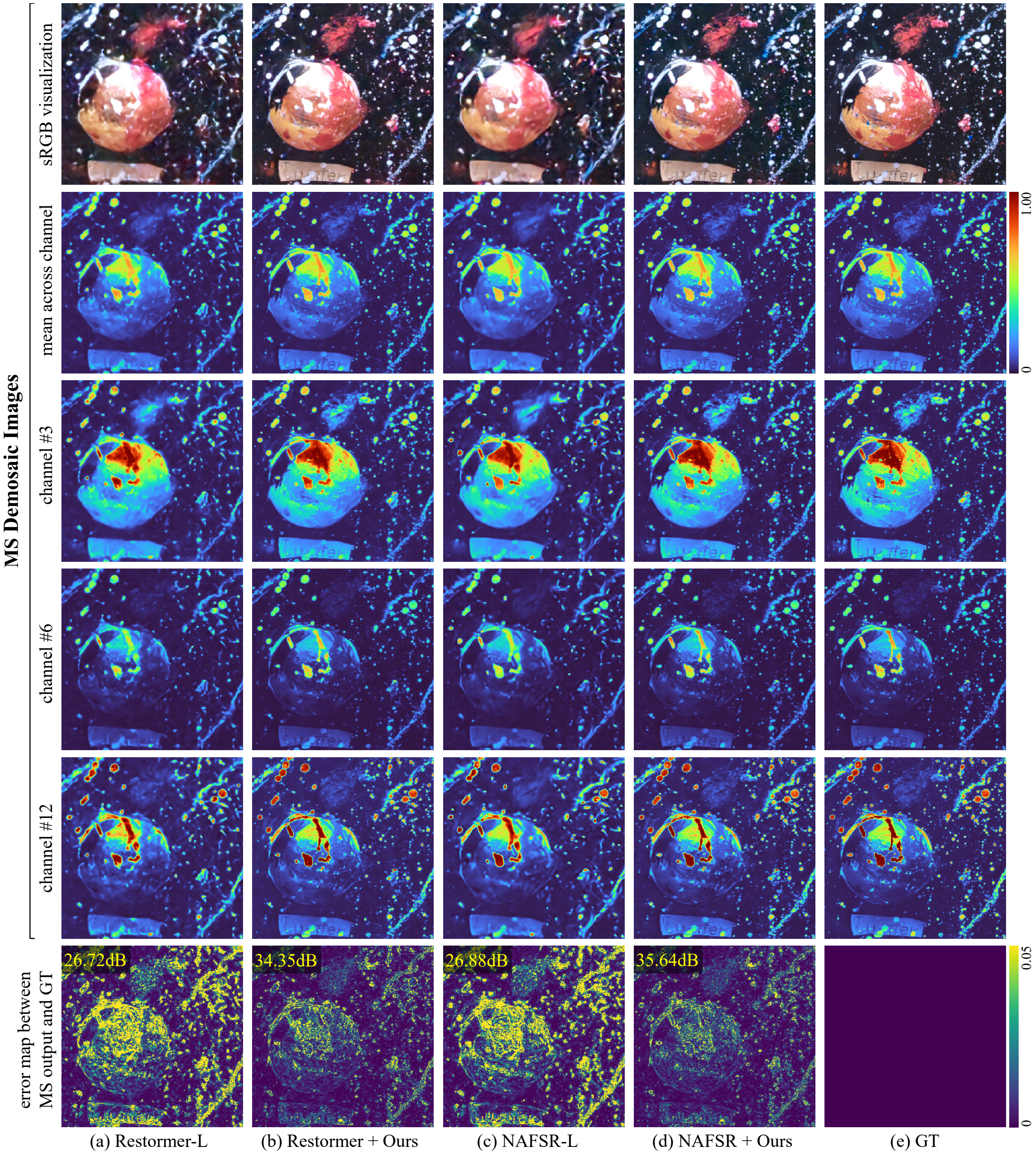
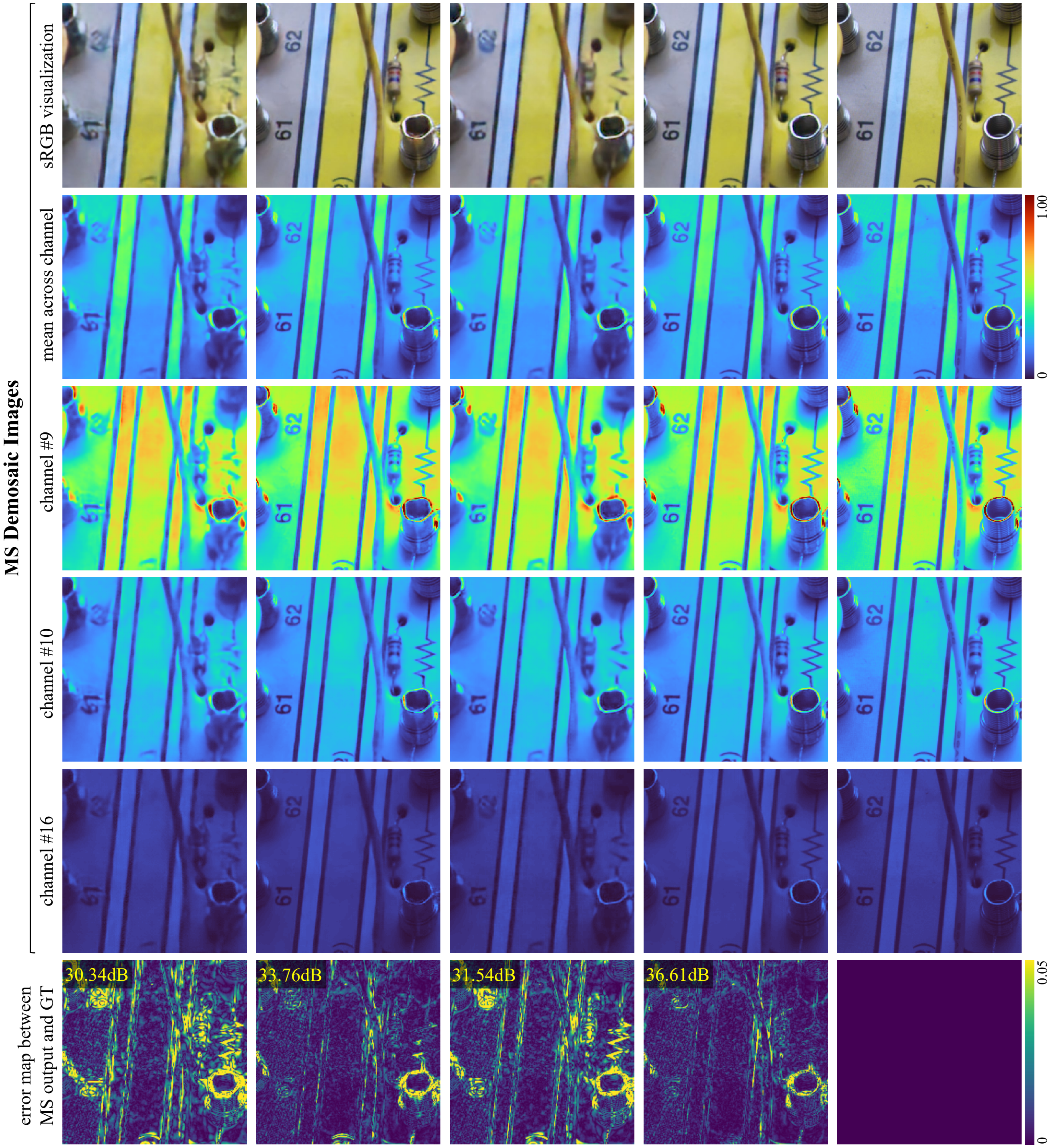


Figure 23. **Qualitative comparison of 4× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 3rd, 6th, and 12th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.



(a) Restormer-L (b) Restormer + Ours (c) NAFSR-L (d) NAFSR + Ours (e) GT

Figure 24. **Qualitative comparison of 4× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 9th, 10th, and 16th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

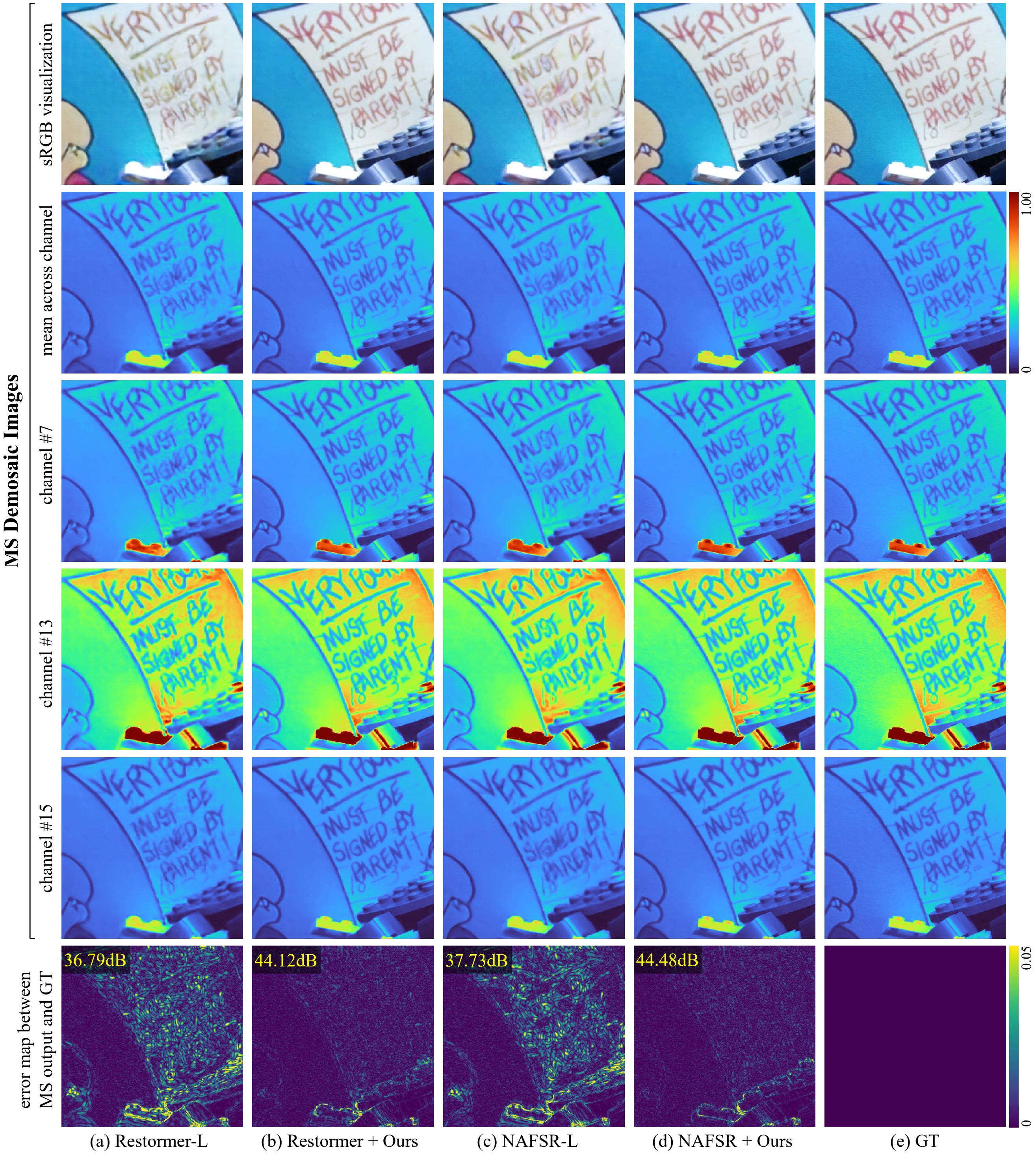


Figure 25. **Qualitative comparison of 4× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 7th, 13th, and 15th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.

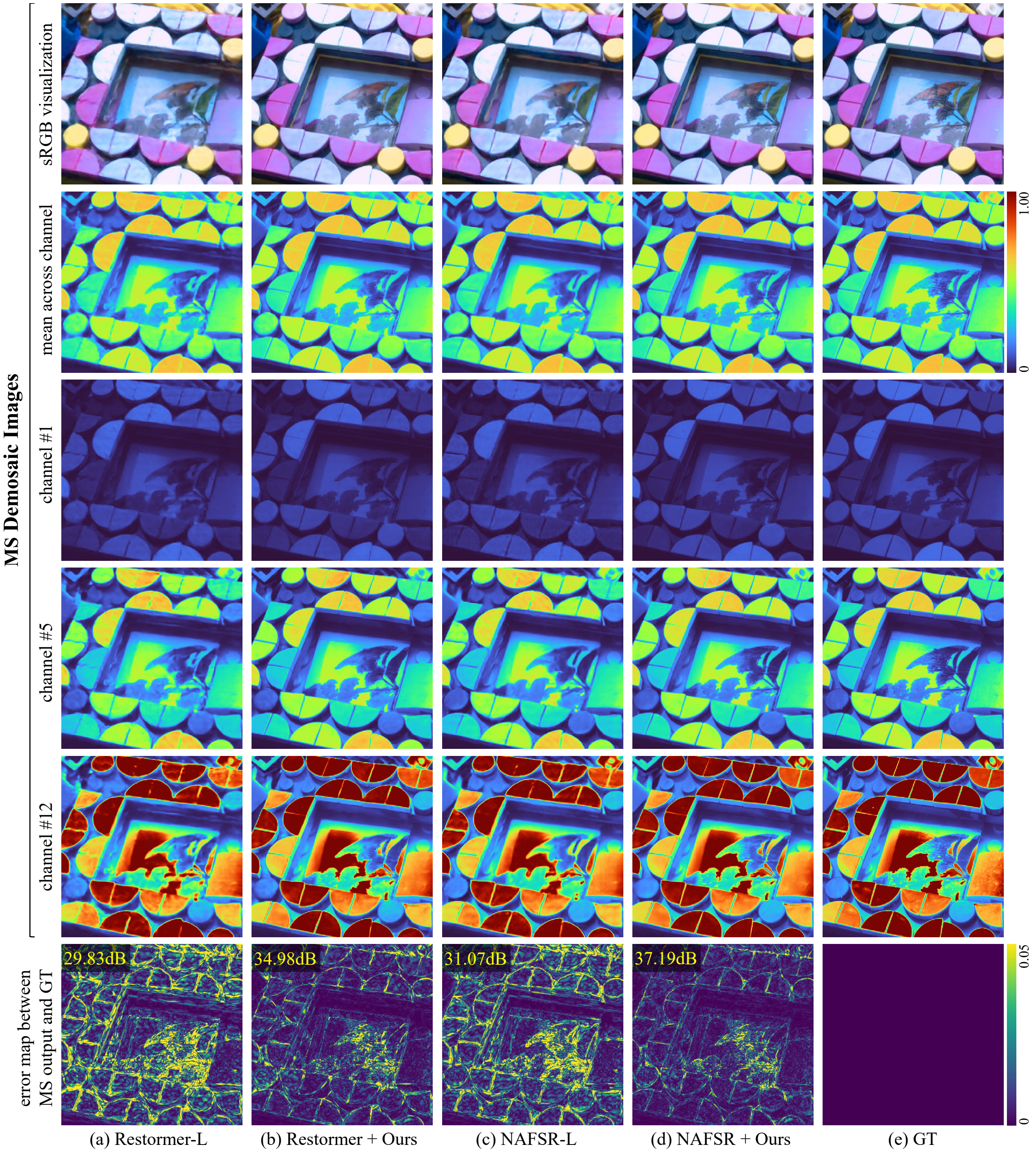


Figure 26. **Qualitative comparison of 4× MS demosaicing results** for a dual-camera scenario featuring MS and RGB sensors with the same spatial resolution but employing asymmetric CFAs. The top row shows the predicted MS demosaics converted to the sRGB color space using the color conversion matrix C (Eq. (5)) and camera metadata, with CIE D65 as the reference white point. The second row presents the MS demosaic output averaged across the channel dimension, while the third to fifth rows display per-channel MS demosaic outputs for the 1st, 5th, and 12th channel indices, respectively. The final row visualizes the error maps between the restored and ground-truth MS demosaiced images.