LEVERAGING MULTISPECTRAL IMAGES FOR CROSS-CAMERA COLOR RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Color restoration, comprising white balance and color correction, is a core stage of the image signal processor (ISP) that maps the raw sensor response into a standard, device-independent color space. Yet with only three broadband RGB channels, disentangling illumination from reflectance is ill-posed, and these operations are tied to a camera-specific raw space that must be recalibrated for every new sensor. Multispectral (MS) sensors sample the spectrum more densely and provide a physical constraint that breaks this underdetermination. Existing methods still underuse this signal, as most apply it only to illuminant estimation, while the few that fuse multispectral image (MSI) with RGB across the whole pipeline train a separate model per camera and cannot transfer to unseen sensors. We present a physics-guided framework that uses a single fixed multispectral camera, treating the MSI as a *camera-agnostic spectral bridge* and exploiting it across the entire color-restoration pipeline. The three-stage framework predicts CSS and white-balance gains, followed by per-image color correction. Crucially, we optimize CSS in the sensor-response domain and use CSS projection to separate sensor dependence from illumination learning. We train the framework once, and it transfers to a new target RGB camera using only the DNG ColorMatrices of the target camera and a small set of MSI–RGB pairs, without retraining. Experimental results show that our framework outperforms existing cross-camera color-constancy and illuminant-estimation baselines and, in the cross-camera setting, approaches the color-restoration accuracy of single-camera MSI methods.

1 INTRODUCTION

Within a camera's image signal processor (ISP), color restoration maps the raw sensor response into a standard, device-independent color space such as CIE XYZ (Karaimer & Brown, 2016); it proceeds in two steps, white balance and then color correction. Both steps rest on separating the scene's illumination from surface reflectance, and this is where the difficulty lies. Compounding it, the whole pipeline is calibrated in one camera's raw space: a model tuned for one sensor does not carry over to another without recalibration or retraining. For a vendor shipping many models, this per-camera cost is a serious obstacle to deployment.

These challenges arise from the image-formation process. Ignoring spatial coordinates, the raw response of color channel $c \in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}$ is

$$\mathrm{RAW}_c = \int R(\lambda)\, E(\lambda)\, S_c(\lambda)\, \mathrm{d}\lambda, \tag{1}$$

where λ is wavelength, $R(\lambda)$ is the surface spectral reflectance, $E(\lambda)$ is the illuminant spectral power distribution (SPD), and $S_c(\lambda)$ is the camera spectral sensitivity (CSS) of channel c. With only three broadband measurements, Eq. 1 is severely underdetermined: different combinations of reflectance and illuminant can produce identical RGB values, a phenomenon known as metamerism, so disentangling illumination and reflectance from RGB alone is ill-posed (Arad & Ben-Shahar, 2016). Moreover, S_c varies from camera to camera, so the same scene induces different raw responses on different sensors, making the problem camera-specific.

Breaking this underdetermination requires richer spectral measurements. Multispectral (MS) sensors provide denser spectral observations that improve illuminant estimation (Khan et al., 2017;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

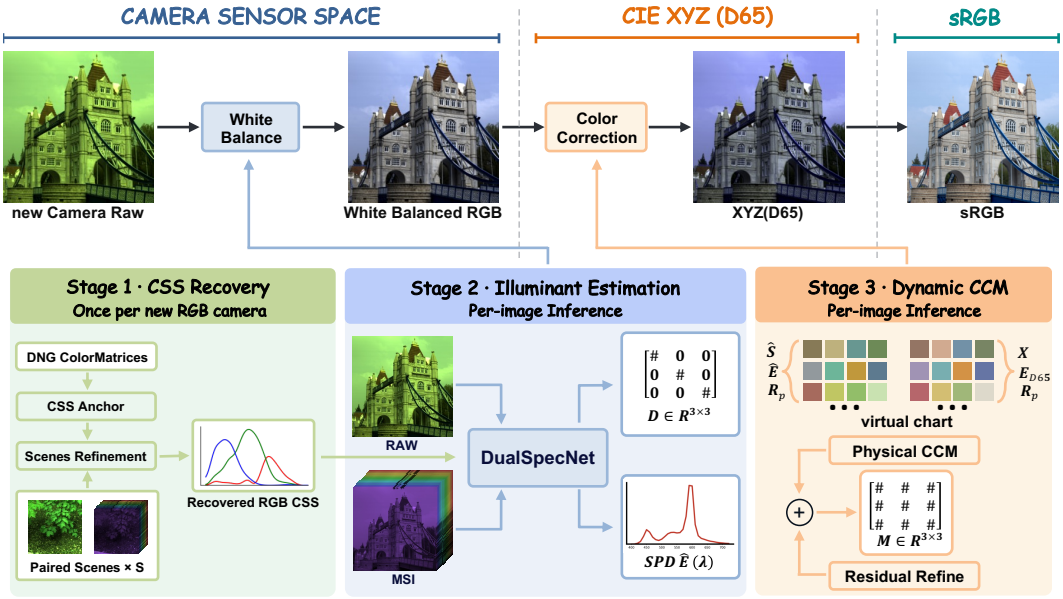

Figure 1: Overview of our cross-camera color-restoration framework. A fixed multispectral camera serves as a camera-agnostic spectral bridge. Stage 1 recovers the target camera's spectral sensitivity (CSS), which Stages 2–3 leverage to estimate the illuminant spectrum, white-balance gains, and per-image CCM, enabling color restoration to generalize across cameras without retraining.

Koskinen et al., 2024; Glatt et al., 2024). Existing MSI-based approaches, however, either stop at illuminant estimation or require camera-specific training when extended to the full color-restoration pipeline (Cogo et al., 2026). In the RGB domain, cross-camera color constancy has been studied extensively (Afifi & Brown, 2019; Afifi et al., 2021; Kim et al., 2025; Hernandez-Juarez et al., 2020; Chang et al., 2025; Li et al., 2026), but these methods remain constrained by the underdetermination and typically stop at white balance, without optimizing the color-correction process. To our knowledge, no existing method uses the multispectral image as a camera-agnostic signal to transfer the complete color restoration across cameras.

A fixed multispectral camera measures the scene spectrum independently of the target RGB camera, so the multispectral image can serve as a *camera-agnostic spectral bridge* that decouples the camera-specific unknowns from the scene. Building on this observation, we present an MSI-based cross-camera color-restoration framework, illustrated in Fig. 1. Rather than learning the mapping end-to-end, we follow the ISP pipeline and separate it into white balance and color correction. The framework is a three-stage cascade consisting of camera spectral sensitivity recovery, joint estimation of the illuminant spectrum and white-balance gains, and per-image color correction. Specifically, we refine CSS to reproduce the target camera's observed responses under shared scene radiance, rather than optimizing curve shape alone. We project a shared spectral illumination representation through the target-camera CSS for white balance, confining sensor dependence to an explicit physical projection.

Our main contributions are as follows:

- We propose a cross-camera color-restoration framework that uses a fixed multispectral camera as a *camera-agnostic spectral bridge* to transfer both white balance and color correction to unseen RGB cameras.
- We develop a target-camera characterization method that recovers CSS through response-domain refinement. The recovered CSS serves as an explicit sensor representation for illuminant estimation and per-image color correction without target-camera retraining.
- Our method outperforms existing cross-camera baselines in white balance and illuminant-spectrum estimation. It also achieves color-restoration accuracy comparable to camera-specific RGB–MSI end-to-end models.

2 Related Work

2.1 White Balance and Color Correction

Illuminant estimation, the first step of white balance, estimates the illuminant color in the camera's raw space (Karaimer & Brown, 2016). Statistical methods rely on global assumptions such as gray-world, white-patch, and gray-edge (Buchsbaum, 1980; Finlayson & Trezzi, 2004; Van De Weijer et al., 2007; Cheng et al., 2014); they are inexpensive but degrade when their assumptions fail. Learning-based methods estimate the illuminant from data-driven features, including log-chrominance histograms, confidence-weighted local estimates, contrastive learning, and attention (Barron, 2015; Barron & Tsai, 2017; Hu et al., 2017; Lo et al., 2021; Kim et al., 2024; Serrano-Lozano et al., 2025). However, these models are typically tied to the raw space of the training camera. Color correction then maps the white-balanced image to a standard space such as CIE XYZ with a linear 3×3 color-correction matrix (CCM). Because the optimal CCM varies with the illuminant, vendors calibrate it at two reference correlated color temperatures (CCTs) and interpolate for other lighting (Karaimer & Brown, 2016; Adobe Inc., 2023). Both steps are defined per camera in the camera's raw space and therefore require recalibration for each new sensor.

2.2 Cross-Camera Color Constancy

Cross-camera color constancy seeks to generalize illuminant estimation across sensor-specific raw spaces. Existing methods map RGB responses to a learned sensor-independent space (Afifi & Brown, 2019), adapt using unlabeled target-camera images (Afifi et al., 2021), or encode pre-calibrated DNG ColorMatrices as camera fingerprints (Kim et al., 2025). Others build this invariance into the training objective (Hernandez-Juarez et al., 2020; Cun et al., 2022), or leverage priors from diffusion models and large language models (Chang et al., 2025; Li et al., 2026). However, they remain subject to the underdetermination in Eq. 1 and generally leave color correction to conventional per-camera CCMs. In contrast, our framework grounds cross-camera transfer in measured multispectral information and applies it to both white balance and color correction.

2.3 Multispectral Imaging for Illuminant Estimation and Color Correction

Multispectral sensors sample spectra more densely than RGB sensors, reducing the metameric ambiguity in Eq. 1. Prior work extends illuminant estimators to multiple channels (Khan et al., 2017; Vidarsson et al., 2026; Jung et al., 2025), exploits the scene-average spectrum (Koskinen et al., 2024; Erba et al., 2024b), or directly predicts the illuminant SPD (Glatt et al., 2024; Habili et al., 2023; Deeb et al., 2018; Oh et al., 2026; Li et al., 2021). Other methods reproduce color from multispectral captures (Li et al., 2025) or fuse RGB and MSI inputs in an end-to-end ISP (Cogo et al., 2026). The latter is trained for a fixed RGB sensor and cannot generalize to unseen cameras. Instead, we treat the MSI as a *camera-agnostic spectral bridge* and enable cross-camera transfer without retraining.

3 Method

3.1 Preliminaries and Problem Setup

Auto white balance and color correction. Under the common single-illuminant assumption, auto white balance (AWB) estimates the global illuminant color $\hat{\mathbf{l}}$ from the image and applies a diagonal white-balance matrix $\mathbf{D} = \mathrm{diag}(\mathbf{g})$ to remove the cast, with gains $\mathbf{g} = (\hat{\ell}_G/\hat{\ell}_R,\, 1,\, \hat{\ell}_G/\hat{\ell}_B)$. A 3×3 color-correction matrix (CCM) $\mathbf{M}$ then maps the white-balanced raw image to CIE XYZ:

$$\mathrm{RAW}_{\mathrm{wb}} = \mathbf{D} \cdot \mathrm{RAW}, \qquad \mathbf{X} = \mathbf{M} \cdot \mathrm{RAW}_{\mathrm{wb}}. \tag{2}$$

The optimal $\mathbf{M}$ depends on both the camera spectral sensitivities and the scene illuminant. DNG pipelines therefore provide two matrix families calibrated under reference illuminants A and D65: ColorMatrices $\mathrm{CM}_{\mathrm{A}}, \mathrm{CM}_{\mathrm{D65}}$ map CIE XYZ to camera raw space, whereas ForwardMatrices $\mathrm{FM}_{\mathrm{A}}, \mathrm{FM}_{\mathrm{D65}}$ map white-balanced raw responses to CIE XYZ. At run time, the ForwardMatrices are interpolated according to the estimated scene CCT:

$$\mathbf{M} = \mathrm{FM}(t) = \mathrm{lerp}(\mathrm{FM}_{\mathrm{A}}, \mathrm{FM}_{\mathrm{D65}}; t), \qquad t \in [0, 1]. \tag{3}$$

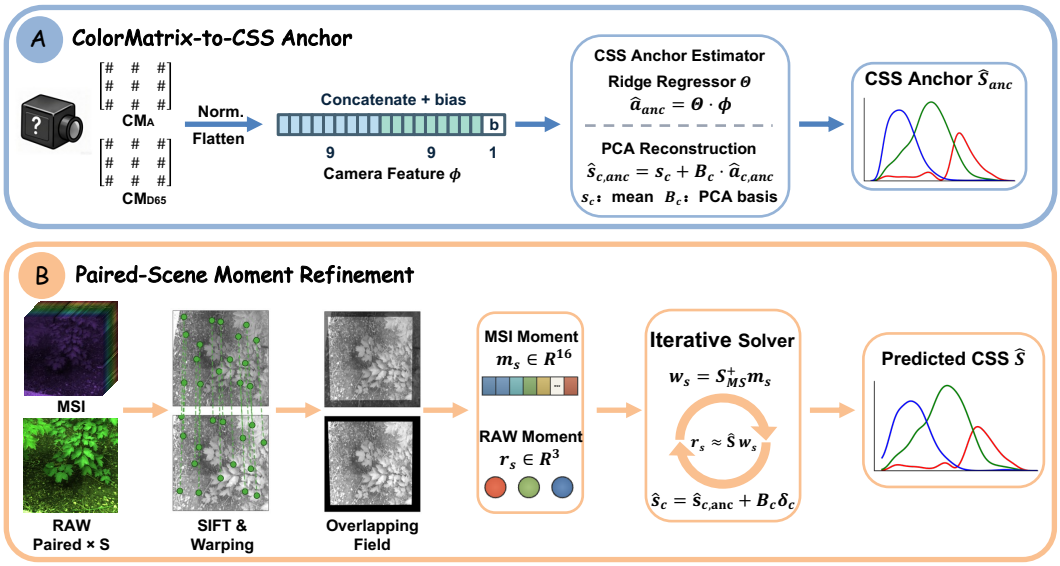

Figure 2: Stage 1: CSS recovery for an unseen RGB camera. (A) A ridge regressor maps two DNG ColorMatrices to a PCA-based CSS anchor. (B) Paired MSI–RAW scene moments refine this anchor in the response domain.

Problem setup. Rather than fitting $\mathbf{D}$ and $\mathbf{M}$ independently, our framework estimates the camera spectral sensitivity $\mathbf{S} = [\mathbf{s}_R, \mathbf{s}_G, \mathbf{s}_B]^\top$, the scene illuminant spectrum $\mathbf{E}$, and the white-balance gains $\mathbf{g}$. The CSS $\mathbf{S}$ is fixed for a given camera and is estimated once as $\hat{\mathbf{S}}$ from its two DNG ColorMatrices and a small set of co-captured MSI–RGB pairs. The recovered $\hat{\mathbf{S}}$ is used to project the predicted white-balance representation into the target camera space and to construct the physical per-image CCM. $\mathbf{E}$ and $\mathbf{g}$ vary with each capture and are estimated at inference from the co-captured MSI and RAW image.

The framework is trained once on a set of source cameras and applied to an unseen target camera after one-time CSS calibration, without retraining. At inference, it takes the target-camera RAW image and an MSI captured by a fixed multispectral module with known sensitivities $\mathbf{S}_{\mathrm{MS}}$, and produces the D65-referenced tristimulus image $\mathbf{X}$. Stage 1 recovers $\hat{\mathbf{S}}$, Stage 2 jointly estimates $\hat{\mathbf{E}}$ and $\mathbf{g}$, and Stage 3 constructs a physical per-image CCM from $\hat{\mathbf{S}}$ and $\hat{\mathbf{E}}$.

3.2 Stage 1: Camera Spectral Sensitivity Recovery

Stage 1 recovers the target CSS $\hat{\mathbf{S}}$ from two complementary sources. We first regress an anchor from the DNG ColorMatrices (Fig. 2A) and then refine its imaging behavior in the response domain using paired MSI–RAW moments (Fig. 2B). Camera spectral sensitivities lie close to a low-dimensional linear subspace (Jiang et al., 2013). To constrain this estimation, we perform PCA separately for each color channel and represent each channel as

$$\mathbf{s}_c \approx \bar{\mathbf{s}}_c + \mathbf{B}_c \cdot \mathbf{a}_c, \qquad c \in \{R, G, B\}, \tag{4}$$

where $\bar{\mathbf{s}}_c$ is the mean sensitivity of channel c, the columns of $\mathbf{B}_c = [\mathbf{b}_{c,1}, \ldots, \mathbf{b}_{c,k}]$ are its top k=10 principal components, and $\mathbf{a}_c \in \mathbb{R}^k$ is the corresponding coefficient vector. CSS recovery therefore reduces to estimating $\mathbf{a} = [\mathbf{a}_R; \mathbf{a}_G; \mathbf{a}_B] \in \mathbb{R}^{3k}$.

ColorMatrix-to-CSS anchor. A ColorMatrix (Sec. 3.1) shows how each illuminant appears in that camera's native RGB domain and thus implicitly encodes the camera's sensing characteristics—an idea similar to CCMNet (Kim et al., 2025). We use the two ColorMatrices to predict the CSS subspace coefficients $\mathbf{a}$. Specifically, we normalize each matrix by its Frobenius norm, vectorize it, and append a bias term:

$$\boldsymbol{\phi} = \left[\, \mathrm{vec}(\mathrm{CM_A})/\|\mathrm{CM_A}\|_F;\;\; \mathrm{vec}(\mathrm{CM_{D65}})/\|\mathrm{CM_{D65}}\|_F;\;\; 1 \,\right] \in \mathbb{R}^{19}. \tag{5}$$

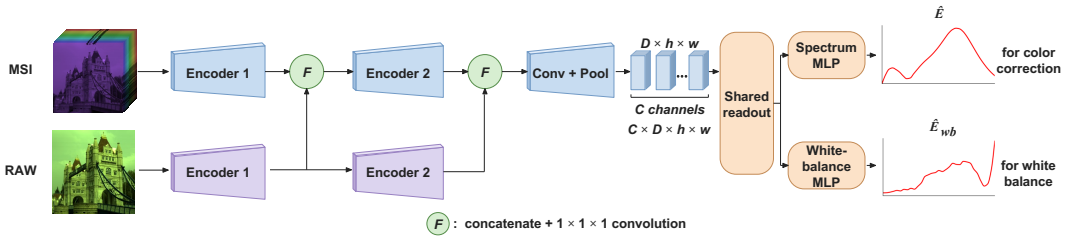

Figure 3: Overview of Stage 2. Two heads estimate the illuminant spectrum and white-balance gains, with the latter obtained through CSS projection.

Given the limited number of source cameras, we use a ridge regressor $\boldsymbol{\Theta} \in \mathbb{R}^{3k\times 19}$ rather than a higher-capacity network. With regularization weight $\gamma{=}10^{-3}$, the objective is

$$\boldsymbol{\Theta} = \arg\min_{\boldsymbol{\Theta}} \sum_{j\in\text{train}} \|\mathbf{a}^{(j)} - \boldsymbol{\Theta}\cdot\boldsymbol{\phi}^{(j)}\|^2 + \gamma\|\boldsymbol{\Theta}\|_F^2, \qquad \hat{\mathbf{a}}_{\text{anc}} = \boldsymbol{\Theta}\cdot\boldsymbol{\phi}. \tag{6}$$

Splitting $\hat{\mathbf{a}}_{\text{anc}}$ per channel and substituting into Eq. 4 gives the anchor $\hat{\mathbf{s}}_{c,\text{anc}} = \bar{\mathbf{s}}_c + \mathbf{B}_c\cdot\hat{\mathbf{a}}_{c,\text{anc}}$.

Paired-scene moment refinement. Spectral-angle error does not directly reflect imaging error because the sensor response integrates the CSS against the scene radiance $R(\lambda)E(\lambda)$, under which local spectral discrepancies may cancel. We therefore refine the anchor in the *response domain*, fitting the target camera's predicted response to its observed raw rather than matching curves directly.

The two sensors observe the same scene radiance $\mathbf{w}(\lambda) = R(\lambda)E(\lambda)$. Discretizing the wavelength domain over Λ and stacking the radiance spectra of N_{pix} pixels into $\mathbf{W}\in\mathbb{R}^{|\Lambda|\times N_{\text{pix}}}$, Eq. 1 becomes

$$\text{MSI} = \mathbf{S}_{\text{MS}}\cdot\mathbf{W}, \qquad \text{RAW} = \mathbf{S}\cdot\mathbf{W}. \tag{7}$$

To account for slight viewpoint differences, we register each RAW image to its corresponding MSI using SIFT, retain their overlapping field of view, and perform the refinement using spatial averages. For calibration scene s, spatial averaging preserves the image-formation model because it is linear:

$$\bar{\mathbf{m}}_s = \mathbf{S}_{\text{MS}}\cdot\bar{\mathbf{w}}_s, \qquad \bar{\mathbf{r}}_s = \mathbf{S}\cdot\bar{\mathbf{w}}_s.$$

We back-project the mean MSI response through the Moore–Penrose pseudoinverse to recover the minimum-norm mean radiance $\tilde{\mathbf{w}}_s = \mathbf{S}_{\text{MS}}^{+}\cdot\bar{\mathbf{m}}_s$, then fit the target camera's predicted mean $\hat{\mathbf{S}}\cdot\tilde{\mathbf{w}}_s$ to the observed mean $\bar{\mathbf{r}}_s$. Writing the recovered sensitivities as the anchor plus a correction, $\hat{\mathbf{s}}_c = \hat{\mathbf{s}}_{c,\text{anc}} + \mathbf{B}_c\,\boldsymbol{\delta}_c$, where $\boldsymbol{\delta}_c\in\mathbb{R}^k$ is the unknown correction, we solve in the response domain

$$\{\boldsymbol{\delta}_c\}^{(n+1)} = \arg\min_{\{\boldsymbol{\delta}_c\}} \sum_s \left\| \mathrm{n}(\bar{\mathbf{r}}_s) - \tfrac{1}{z_s^{(n)}}\,\hat{\mathbf{S}}\cdot\tilde{\mathbf{w}}_s \right\|^2 + \lambda_{\text{sm}}\sum_c \|\mathbf{L}\cdot\hat{\mathbf{s}}_c\|^2, \tag{8}$$

Here, $\mathrm{n}(\mathbf{x}) = \mathbf{x}/\sum_c x_c$ denotes channel-sum normalization. The predicted-response normalization factor $z_s^{(n)} = \sum_c \big(\hat{\mathbf{S}}^{(n)}\tilde{\mathbf{w}}_s\big)_c$ is evaluated at the previous iterate and held fixed during the current update, making each update a regularized linear least-squares problem. This channel-sum normalization removes the unknown per-scene scale. $\mathbf{L}$ computes second-order differences along the sampled sensitivity curve and thereby imposes spectral smoothness. In practice, we set $\lambda_{\text{sm}} = 3\times 10^{-4}$.

3.3 Stage 2: Illuminant Spectrum and White-Balance Gains

Given the CSS $\hat{\mathbf{S}}$ recovered by Stage 1, Stage 2 jointly estimates the illuminant spectrum $\hat{\mathbf{E}}$ and the target-camera white-balance gains $\mathbf{g}$ from the co-captured MSI and RAW images. The overall architecture is illustrated in Fig. 3.

Spectral-preserving backbone. We modify the spectrum-aware estimator of Oh et al. (2026) in two ways. First, we preserve the spectral axis throughout the backbone rather than collapsing it early, thereby retaining wavelength-specific illumination cues. Second, we replace the fully connected layer over the flattened feature map with a lightweight spectral readout head. This head applies mean, standard-deviation, and maximum pooling over the spatial dimensions, followed by convolution along the spectral axis. Since illumination is a global scene property, spatial pooling supports arbitrary input resolutions and reduces sensitivity to misalignment between the MSI and RAW images. These changes also reduce the parameter count to approximately $1/24$ of the baseline.

Two-branch MSI–RAW design. The estimator has two branches. The MSI branch encodes the 16-band multispectral image and is camera-agnostic, whereas the RAW branch encodes the 3-channel target RAW. The two branches are fused progressively. The RAW contribution at each fusion point is *zero-initialized*, so training starts from a pure-MSI estimator and gradually incorporates the RAW signal.

Dual heads with CSS-projected white balance. Two heads each regress a spectral output from the fused features using a Softplus activation and ℓ_2 normalization. The spectrum head predicts the illuminant spectrum $\hat{\mathbf{E}}$, while the white-balance head predicts a task-specific spectral representation $\hat{\mathbf{E}}_{\mathrm{wb}}$. Integrating the latter with the recovered CSS gives the camera-space white point $\hat{\mathbf{l}}$ and the corresponding gains:

$$\hat{\mathbf{l}} = \hat{\mathbf{S}} \cdot \hat{\mathbf{E}}_{\mathrm{wb}}, \qquad \mathbf{g} = (\hat{\ell}_G/\hat{\ell}_R,\ 1,\ \hat{\ell}_G/\hat{\ell}_B). \tag{9}$$

The two heads are trained with angular losses on $\hat{\mathbf{E}}$ and $\hat{\mathbf{l}}$, respectively. The spectrum head provides high-dimensional spectral supervision that encourages the shared backbone to separate scene illumination from surface reflectance, yielding $\hat{\mathbf{E}}$ for Stage 3 and a more informative representation for white-point estimation. On the other hand, since errors in the estimated illuminant spectrum do not necessarily reflect white-point errors, we introduce a separate white-balance head to learn which bands of the illuminant matter most for the white point. By accounting for sensor dependence explicitly through the CSS projection, this formulation encourages $\hat{\mathbf{E}}_{\mathrm{wb}}$ to represent illumination in a shared spectral domain, thereby enabling cross-camera white-balance estimation without retraining.

3.4 Stage 3: Physical Per-Image CCM with Residual Refinement

The final stage constructs a scene-adaptive CCM that maps the white-balanced RAW image to CIE XYZ. Conventional CCMs are calibrated from a photographed color chart under two reference illuminants. Instead, we simulate the same calibration process using a virtual chart rendered under the scene illuminant estimated by Stage 2. This construction is enabled by the full illuminant spectrum and CSS available in our multispectral setting. Consequently, at deployment, the physical CCM can be derived without paired RAW–XYZ ground truth or target-camera-specific retraining.

Virtual-chart calibration. Concretely, given the recovered sensitivity $\hat{\mathbf{S}}$ and the estimated illuminant spectrum $\hat{\mathbf{E}}$, we select a fixed reflectance library $\{\mathbf{R}_p\}_{p=1}^{N_R}$ from a reflectance dataset to serve as the virtual chart. For each reflectance, we render a white-balanced camera response $\boldsymbol{\rho}_p$ under the estimated scene illuminant and a reference CIE XYZ tristimulus $\mathbf{x}_p$ under D65:

$$\boldsymbol{\rho}_p = \mathbf{D} \cdot \hat{\mathbf{S}} \cdot (\mathbf{R}_p \odot \hat{\mathbf{E}}), \qquad \mathbf{x}_p = \kappa\, \bar{\mathbf{X}} \cdot (\mathbf{R}_p \odot \mathbf{E}_{\mathrm{D65}}), \tag{10}$$

where $\mathbf{D}$ is the diagonal white-balance matrix, $\odot$ denotes element-wise multiplication along wavelength, $\bar{\mathbf{X}} \in \mathbb{R}^{3\times|\Lambda|}$ is the matrix of CIE 1931 color-matching functions, and $\kappa = 1/\sum_\lambda \bar{y}(\lambda) E_{\mathrm{D65}}(\lambda)$ normalizes a unit white reflector to $Y{=}1$.

Let $\mathbf{P} = [\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_{N_R}]$ and $\mathbf{T} = [\mathbf{x}_1, \ldots, \mathbf{x}_{N_R}]$. Assuming that $\mathbf{P}$ has full row rank, the least-squares physical CCM is

$$\mathbf{M}_{\mathrm{phys}} = \mathbf{T} \cdot \mathbf{P}^\top (\mathbf{P} \cdot \mathbf{P}^\top)^{-1}. \tag{11}$$

Residual refinement. The physical CCM is optimal only for the chosen reflectance library and also inherits errors from the estimated $\hat{\mathbf{S}}$ and $\hat{\mathbf{E}}$. We therefore treat it as a physical anchor and learn an additive residual:

$$\mathbf{M} = \mathbf{M}_{\mathrm{phys}} + \Delta\mathbf{M}_\psi\big(\mathrm{RAW}_{\mathrm{wb}}, \hat{\mathbf{S}}, \mathbf{M}_{\mathrm{phys}}, \hat{\mathbf{l}}\big) \tag{12}$$

where $\Delta\mathbf{M}_\psi$ predicts a single 3×3 correction for each image. The network encodes $\mathrm{RAW}_{\mathrm{wb}}$ with a CNN and encodes $\hat{\mathbf{S}}$, $\mathbf{M}_{\mathrm{phys}}$, and $\hat{\mathbf{l}}$ with separate MLPs. We zero-initialize its output layer so that refinement starts exactly from $\mathbf{M}_{\mathrm{phys}}$. The final D65-referenced CIE XYZ output is $\mathbf{X}_{\mathrm{out}} = \mathbf{M} \cdot \mathrm{RAW}_{\mathrm{wb}}$. We train the residual network using the ΔE_{00} color difference (Sharma et al., 2005). To make the residual refinement robust to upstream estimation errors, we perturb $\hat{\mathbf{S}}$ and $\hat{\mathbf{E}}$ during training with random low-frequency sinusoidal distortions of fixed angular magnitude and recompute $\mathbf{M}_{\mathrm{phys}}$ online.

Table 1: Comparison with RGB-based cross-camera methods on the 14 unseen cameras. All methods use the interpolated DNG ForwardMatrix in Eq. 3 for color correction. For our framework, we additionally report results with the proposed scene-adaptive CCM. Best results are shown in bold and second-best results are underlined.

Method	White-point angular error (°)↓					ΔE_{00} ↓				
	Mean	Med.	Tri.	B25%	W25%	Mean	Med.	Tri.	B25%	W25%
Gray-World (Buchsbaum, 1980)	3.96	3.75	3.73	1.03	7.23	6.85	7.00	6.78	2.62	10.92
Gray-Edge (Van De Weijer et al., 2007)	3.53	3.10	3.13	1.05	6.87	6.45	6.31	6.30	2.91	10.54
PCA (Cheng et al., 2014)	3.32	3.03	3.02	0.81	6.46	6.18	5.87	5.84	2.61	10.25
SIIE (Afifi & Brown, 2019)	1.83	1.39	1.50	0.46	3.90	4.10	3.87	3.91	1.95	6.66
C5 (Afifi et al., 2021)	<u>1.45</u>	<u>1.19</u>	<u>1.25</u>	<u>0.37</u>	<u>2.97</u>	3.93	3.73	3.77	1.89	6.34
DMCC (Yue & Wei, 2024)	2.02	1.61	1.68	0.54	4.24	4.64	4.35	4.42	2.05	7.70
CCMNet (Kim et al., 2025)	1.61	1.23	1.29	0.40	3.46	4.03	3.74	3.76	<u>1.76</u>	6.82
GCC (Chang et al., 2025)	2.09	1.94	1.97	0.82	3.63	4.91	4.99	4.89	2.52	7.29
VLM-CC (Li et al., 2026)	1.49	1.30	1.35	0.48	**2.81**	<u>3.70</u>	<u>3.54</u>	<u>3.58</u>	1.88	<u>5.80</u>
Ours (interpolated CCM)	**1.35**	**0.95**	**1.02**	**0.32**	3.08	3.86	3.63	3.68	1.82	6.27
Ours (scene-adaptive CCM)	**1.35**	**0.95**	**1.02**	**0.32**	3.08	**2.38**	**1.87**	**2.00**	**0.86**	**4.74**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data. As no public dataset provides paired MSI–RAW observations across multiple cameras with color ground truth, we construct a synthetic benchmark following Cogo et al. (2026). We split 765 reflectance scenes (Du et al., 2025) into $612/76/77$ for training/validation/test. Each scene is paired with 20 spectra sampled from disjoint pools of $1286/161/161$ measured illuminants (Royer, 2023; Barnard et al., 2002). For each pair, we render a 16-band MSI using the multispectral camera sensitivities from NTIRE 2022 (Arad et al., 2022), a RAW image using a randomly selected RGB-camera sensitivity (Dyer et al., 2019), and a D65-referenced CIE XYZ target. The corresponding DNG ColorMatrices are obtained from Adobe profiles. To simulate the different spatial resolutions of RGB and multispectral cameras, we render the RAW and MSI images at 512×512 and 256×256 respectively. To simulate cross-sensor misalignment, we apply a random Zurich homography (Ignatov et al., 2020) to each aligned RAW–XYZ pair. Training uses 36 cameras from Canon, Fujifilm, and Nikon, while validation and testing use 14 cameras from the held-out manufacturers Sony, Hasselblad, and Panasonic. Experiments on real-world data are presented in Appendix G.

Implementation Details. Stage 1 is solved in closed form using 60 MSI–RAW pairs per target camera. Stages 2 and 3 are trained separately for 300 epochs using AdamW with a weight decay of 3×10^{-4} and a cosine learning-rate schedule. Stage 2 uses a learning rate of 10^{-3} and a batch size of 48, with spectral and white-point angular losses weighted 1:2. Stage 3 uses a learning rate of 10^{-4} and a batch size of 16, and is trained with the ΔE_{00} loss (Sharma et al., 2005).

4.2 COLOR-RESTORATION COMPARISON

Comparison with RGB cross-camera methods. Since the RGB-based baselines estimate white balance only, we apply the interpolated DNG ForwardMatrix in Eq. 3 for color correction. We retrain all baselines with available training code on our synthetic dataset and use the official weights for GCC and VLM-CC. As shown in Tab. 1, our framework achieves the lowest mean white-balance error of $1.35°$ and the lowest mean ΔE_{00} of 2.38. These gains arise from two complementary components: improved white balance reduces the residual color cast, while the scene-adaptive CCM uses the MSI-derived illuminant spectrum to adapt color correction to each image beyond conventional DNG interpolation.

Comparison with single-camera methods. We compare our framework with single-camera methods on the Sony ILCE-7RM2. Following Cogo et al. (2026), we adapt FC4 and ConvMean to accept MSI inputs. For illuminant-estimation methods, we apply the interpolated DNG ForwardMatrix after white balancing to obtain the final CIE XYZ output. All learning-based baselines are trained specifically on the test camera, whereas our model is transferred from the 36 source

Table 2: Comparison with single-camera baselines on the Sony ILCE-7RM2. Learning-based baselines are trained specifically on the test camera, whereas our framework transfers from the source cameras without retraining. Best results are shown in bold and second-best results are underlined.

Setting	Method	Input	$\Delta E_{00} \downarrow$				
			Mean	Med.	Tri.	B25%	W25%
Single-camera	FFCC (Barron & Tsai, 2017)	RGB	3.96	3.73	3.79	1.87	6.45
	SqueezeNet-FC4 (Hu et al., 2017)	RGB	3.69	3.50	3.51	1.88	5.89
	ConvMean (Gong, 2020)	RGB	4.01	3.67	3.75	1.73	6.93
	QU (Bianco & Cusano, 2019)	RGB	6.29	6.06	6.01	2.59	10.25
	QU+ft (Bianco & Cusano, 2019)	RGB	5.78	5.74	5.69	2.28	9.23
	SpectralFC4 (Hu et al., 2017; Cogo et al., 2026)	MSI	4.26	4.04	4.08	2.11	6.82
	SpectralConvMean (Gong, 2020; Cogo et al., 2026)	MSI	4.20	3.98	3.99	1.91	7.01
	MSCC (LPIENet) (Cogo et al., 2026)	RGB+MSI	2.56	2.31	<u>2.36</u>	1.53	**4.07**
	MSCC (cmKAN-light) (Cogo et al., 2026)	RGB+MSI	**2.31**	<u>1.93</u>	**2.02**	<u>1.05</u>	<u>4.17</u>
Cross-camera	Ours (interpolated CCM)	RGB+MSI	3.76	3.53	3.58	1.84	6.07
	Ours (scene-adaptive CCM)	RGB+MSI	<u>2.42</u>	**1.90**	**2.02**	**0.84**	4.88

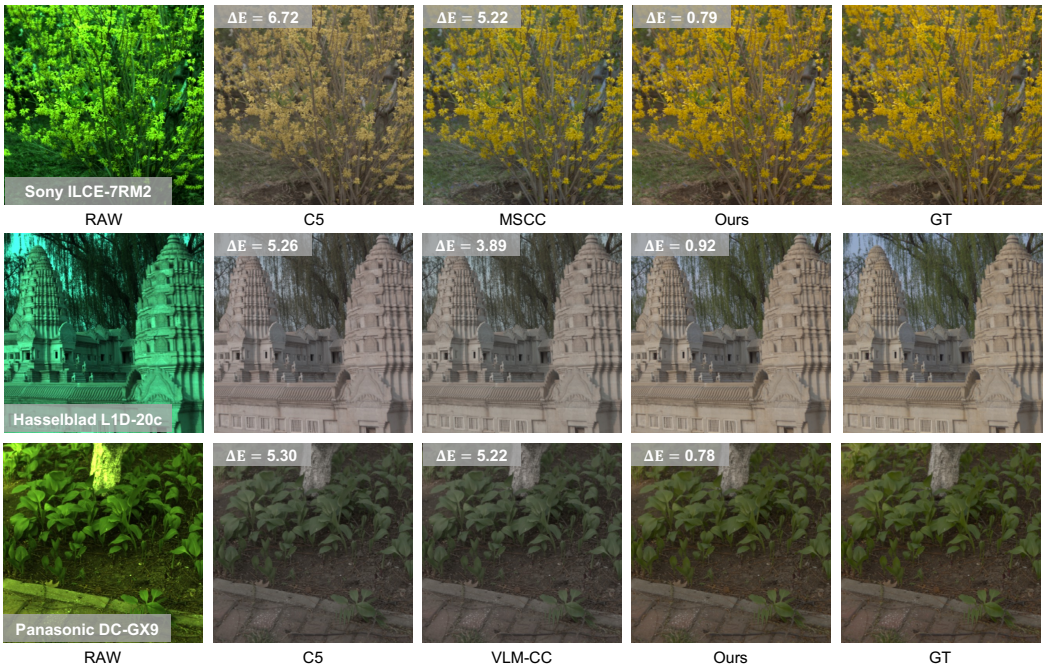

Figure 4: Qualitative color-restoration comparisons. Top: comparison with C5 and the camera-specific MSCC on the Sony ILCE-7RM2 single-camera benchmark. Bottom: cross-camera comparison with C5 and VLM-CC on two unseen cameras.

cameras without retraining. Under this cross-camera setting, our scene-adaptive CCM achieves a mean ΔE_{00} of 2.42, comparable to the single-camera end-to-end models MSCC (LPIENet) at 2.56 and MSCC (cmKAN-light) at 2.31. It further achieves the best median (1.90) and best-quartile error (0.84), while tying the best trimean (2.02). These results demonstrate that the camera-agnostic spectral bridge enables retraining-free transfer to unseen RGB cameras while maintaining performance comparable to camera-specific end-to-end models.

Figure 4 provides qualitative comparisons under both evaluation protocols. On the Sony ILCE-7RM2 example, C5 and our method transfer directly from the source cameras, whereas MSCC is trained specifically on Sony data. On the Hasselblad L1D-20c and Panasonic DC-GX9 examples, all methods operate in the cross-camera setting. Even in these complex natural scenes with diverse surface reflectances, our results closely match the ground truth and largely remove the color casts visible in the competing methods.

Table 3: Stage-wise ablations of the proposed framework. In (a), the configuration without a CM anchor is initialized with the training-camera PCA mean. Panel (b) isolates the backbone and input fusion, (c) fixes the backbone and ablates the prediction heads, and (d) evaluates residual CCM refinement. Best results are shown in bold.

(a) CSS calibration

CM anchor	Paired moments	Smooth.	CSS angle↓	Imaging error↓
✓			6.94	2.22
✓	✓		32.10	5.31
✓	✓	✓	**4.34**	**0.157**

(b) Backbone and input fusion

MSI backbone	RAW branch	SPD mean↓	AWB mean↓
Spectrum		5.28	1.93
Spectral-pres.		5.08	1.71
Spectral-pres.	✓	4.91	1.51

(c) Prediction heads and CSS projection

Spectrum head	White-balance head	SPD mean↓	AWB mean↓
✓		4.91	1.51
	✓ (CSS-projected)	–	1.79
✓	✓ (direct)	5.52	2.23
✓	✓ (CSS-projected)	**4.71**	**1.35**

(d) CCM construction and refinement

CCM initialization	Residual	Error injection	$\Delta E_{00}\downarrow$
Interpolated			3.86
Physical			2.59
Physical	✓		2.59
Physical	✓	✓	**2.38**

Illuminant-spectrum estimation. Stage 2 achieves a mean SPD angular error of $4.71°$, outperforming existing multispectral estimators (Erba et al., 2024a; Robles-Kelly & Wei, 2018; Glatt et al., 2024) and the Spectrum backbone it builds on (Oh et al., 2026); relative to the latter, our improved model reduces the parameter count from 16.3M to 0.68M while improving accuracy. Additional results are provided in Appendix D.

4.3 Ablation Study

Tab. 3 evaluates the main components of the framework. In panel (a), smooth response-domain refinement reduces the CSS/imaging angular errors from $6.94°/2.22°$ to $4.34°/0.157°$, confirming that it better preserves the camera's image-formation behavior. Panel (b) shows that preserving the spectral axis improves both SPD and AWB estimation, while incorporating RAW features further reduces the errors to $4.91°/1.51°$. Panel (c) shows that the two heads play complementary roles. The spectrum head provides high-dimensional supervision that helps the network separate scene illumination from surface reflectance. The CSS-projected white-balance head further encourages the network to learn camera-independent illumination features. Together, they improve the SPD/AWB errors from $4.91°/1.51°$ to $4.71°/1.35°$. In contrast, direct white-point regression requires the backbone to encode camera-specific responses and degrades both errors to $5.52°/2.23°$. In panel (d), the physical CCM lowers ΔE_{00} from 3.86 to 2.59, and error-injected residual refinement further reduces it to 2.38. As an additional control, a directly learned CCM that predicts the complete matrix without a physical initialization obtains 2.565, compared with 2.395 for our physical anchor plus error-injected residual. Our method is better on all 14 test cameras, supporting the benefit of combining physical initialization with robust residual learning; full results are given in Appendix E.

5 Conclusion

We presented a cross-camera color-restoration framework that uses a multispectral image as a *camera-agnostic spectral bridge*. The central idea is to represent each target camera explicitly by its CSS, estimated through a one-time calibration that matches sensor responses rather than curve shape alone. The network predicts a camera-agnostic illumination representation in a shared spectral space and projects it through the target-camera CSS to recover the camera-space white point, avoiding the direct learning of sensor-specific RGB mappings. Under the same physical factorization, the recovered CSS and illuminant spectrum determine a per-image CCM, enabling both white balance and color correction to transfer across cameras without retraining. Across 14 held-out cameras, our method achieves a mean white-balance error of $1.35°$ and a mean ΔE_{00} of 2.38, outperforming cross-camera baselines and approaching the performance of camera-specific RGB–MSI models.

AI USE STATEMENT

In this work, we used generative AI tools to assist with designing and refining the research methodology and experiments, implementing parts of the proposed method, improving the English-language presentation of parts of the manuscript, and interpreting experimental results. We also used generative AI tools to suggest experimental parameters; create, edit, and debug software code; draft and revise parts of the manuscript; improve its readability and structure; search for, identify, summarize, and analyze relevant literature; and format references. We did not use generative AI tools for other tasks requiring mandatory disclosure.

All AI-assisted text and translations were reviewed and revised by the authors. AI-assisted literature summaries and references were checked against the original publications. Suggestions concerning the methodology, experimental design, parameters, and interpretation of results were independently evaluated and verified by the authors. All AI-assisted code was manually inspected, tested, and validated through the experiments reported in this paper. The authors take responsibility for the final content of this work, including its text, claims, code, results, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce our results throughout the paper and appendix. The method section and Section 4.1 specify the three-stage framework, training objectives, benchmark construction, camera split, optimization settings, and hyperparameters. Appendix A further documents the exact data splits, target-camera calibration set, spectral resampling and integration procedure, and color-correction evaluation protocol; the remaining appendices report calibration stability, sensitivity analyses, oracle studies, real-world evaluation, and per-camera results. We will provide the implementation code in the supplementary material to facilitate independent reproduction of the data-generation and evaluation pipelines.

REFERENCES

Adobe Inc. *Digital Negative (DNG) Specification, Version 1.7.1.0*. Adobe Inc., September 2023. URL https://helpx.adobe.com/content/dam/help/en/camera-raw/digital-negative/jcr_content/root/content/flex/items/position/position-par/download_section_733958301/download-1/DNG_Spec_1_7_1_0.pdf.

Mahmoud Afifi and Michael S Brown. Sensor-independent illumination estimation for dnn models. *arXiv preprint arXiv:1912.06888*, 2019.

Mahmoud Afifi, Jonathan T Barron, Chloe LeGendre, Yun-Ta Tsai, and Francois Bleibel. Cross-camera convolutional color constancy. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1961–1970. IEEE, 2021.

Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In *European conference on computer vision*, pp. 19–34. Springer, 2016.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 863–881, 2022.

Kobus Barnard, Lindsay Martin, Brian Funt, and Adam Coath. A data set for color research. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 27(3):147–151, 2002.

Jonathan T Barron. Convolutional color constancy. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 379–387, 2015.

Jonathan T Barron and Yun-Ta Tsai. Fast fourier color constancy. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 6950–6958. IEEE, 2017.

Simone Bianco and Claudio Cusano. Quasi-unsupervised color constancy. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12204–12213. IEEE, 2019.

Gershon Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin institute*, 310(1):1–26, 1980.

Chen-Wei Chang, Cheng-De Fan, Chia-Che Chang, Yi-Chen Lo, Yu-Chee Tseng, Jiun-Long Huang, and Yu-Lun Liu. Gcc: Generative color constancy via diffusing a color checker. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10868–10878. IEEE, 2025.

Dongliang Cheng, Dilip K Prasad, and Michael S Brown. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *Journal of the Optical Society of America A*, 31(5):1049–1058, 2014.

Luca Cogo, Marco Buzzelli, Simone Bianco, Javier Vazquez-Corral, and Raimondo Schettini. Leveraging multispectral sensors for color correction in mobile cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12438–12447, 2026.

Xiaodong Cun, Zhendong Wang, Chi-Man Pun, Jianzhuang Liu, Wengang Zhou, Xu Jia, and Houqiang Li. Learning enriched illuminants for cross and single sensor color constancy. *arXiv preprint arXiv:2203.11068*, 2022.

Rada Deeb, Joost Van de Weijer, Damien Muselet, Mathieu Hebert, and Alain Tremeau. Deep spectral reflectance and illuminant estimation from self-interreflections. *Journal of the Optical Society of America A*, 36(1):105–114, 2018.

Zhuoran Du, Shaodi You, Cheng Cheng, and Shikui Wei. Automatic spectral calibration of hyperspectral images: Method, dataset and benchmark. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28081–28090. IEEE, 2025.

Scott Dyer, Alexander Forsythe, Jonathon Irons, Thomas Mansencal, and Miaoqi Zhu. RAW to ACES utility data, 2019. URL `https://doi.org/10.5281/zenodo.3372171`.

Ilaria Erba, Marco Buzzelli, and Raimondo Schettini. Rgb color constancy using multispectral pixel information. *Journal of the Optical Society of America A*, 41(2):185–194, 2024a.

Ilaria Erba, Marco Buzzelli, Jean-Baptiste Thomas, Jon Yngve Hardeberg, and Raimondo Schettini. Improving rgb illuminant estimation exploiting spectral average radiance. *Journal of the Optical Society of America A*, 41(3):516–526, 2024b.

Graham D Finlayson and Elisabetta Trezzi. Shades of gray and colour constancy. In *Color and imaging conference*, volume 12, pp. 37–41. Society of Imaging Science and Technology, 2004.

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 *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4332–4342. IEEE, 2024.

Han Gong. Convolutional mean: A simple convolutional neural network for illuminant estimation. *arXiv preprint arXiv:2001.04911*, 2020.

Nariman Habili, Jeremy Oorloff, and Lars Petersson. Automatic illumination spectrum recovery. *arXiv preprint arXiv:2305.19538*, 2023.

Daniel Hernandez-Juarez, Sarah Parisot, Benjamin Busam, Ales Leonardis, Gregory Slabaugh, and Steven McDonagh. A multi-hypothesis approach to color constancy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2270–2280, 2020.

Yuanming Hu, Baoyuan Wang, and Stephen Lin. Fcˆ 4: Fully convolutional color constancy with confidence-weighted pooling. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 330–339. IEEE, 2017.

Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2275–2285. IEEE, 2020.

Jun Jiang, Dengyu Liu, Jinwei Gu, and Sabine Süsstrunk. What is the space of spectral sensitivity functions for digital color cameras? In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 168–179. IEEE, 2013.

Geunho Jung, Semin Kim, and Sangwook Yoo. Comparative study of multispectral image-based auto white balance with optimized conditions. *Journal of Spectroscopy*, 2025(1):7775119, 2025.

Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In *European Conference on Computer Vision*, pp. 429–444. Springer, 2016.

Haris Ahmad Khan, Jean-Baptiste Thomas, Jon Yngve Hardeberg, and Olivier Laligant. Illuminant estimation in multispectral imaging. *Journal of the Optical Society of America A*, 34(7):1085–1098, 2017.

Dongyoung Kim, Jinwoo Kim, Junsang Yu, and Seon Joo Kim. Attentive illumination decomposition model for multi-illuminant white balancing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25512–25521. IEEE, 2024.

Dongyoung Kim, Mahmoud Afifi, Dongyun Kim, Michael S Brown, and Seon Joo Kim. Ccmnet: Leveraging calibrated color correction matrices for cross-camera color constancy. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6198–6208. IEEE, 2025.

Samu Koskinen, Erman Acar, and Joni-Kristian Kämäräinen. Single pixel spectral color constancy: S. koskinen et al. *International Journal of Computer Vision*, 132(2):287–299, 2024.

Jiacheng Li, Chang Chen, Xue Hu, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Multi-spectral image color reproduction. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8400–8409. IEEE, 2025.

Shuwei Li, Lei Tan, and Robby T Tan. White-balance first, adjust later: Cross-camera color constancy via vision-language evaluation. *arXiv preprint arXiv:2605.19613*, 2026.

Yuqi Li, Qiang Fu, and Wolfgang Heidrich. Multispectral illumination estimation using deep unrolling network. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2652–2661. IEEE, 2021.

Yi-Chen Lo, Chia-Che Chang, Hsuan-Chao Chiu, Yu-Hao Huang, Chia-Ping Chen, Yu-Lin Chang, and Kevin Jou. Clcc: Contrastive learning for color constancy. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8049–8059. IEEE, 2021.

Hyejin Oh, Woo-Shik Kim, Sangyoon Lee, YungKyung Park, and Je-Won Kang. Spectrum aware illumination estimation using multispectral image. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Antonio Robles-Kelly and Ran Wei. A convolutional neural network for pixelwise illuminant recovery in colour and spectral images. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 109–114. IEEE, 2018.

Michael Royer. Real light source SPDs and color data for use in research. Figshare, December 2023. URL `https://doi.org/10.6084/m9.figshare.12947240.v2`. Version 2.

David Serrano-Lozano, Aditya Arora, Luis Herranz, Konstantinos G Derpanis, Michael S Brown, and Javier Vazquez-Corral. Revisiting image fusion for multi-illuminant white-balance correction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8275–8284. IEEE, 2025.

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

Joost Van De Weijer, Theo Gevers, and Arjan Gijsenij. Edge-based color constancy. *IEEE Transactions on image processing*, 16(9):2207–2214, 2007.

G Dofri Vidarsson, Liying Lu, and Sabine Süsstrunk. Color constancy in hyperspectral imaging via reduced spectral spaces. *arXiv preprint arXiv:2605.13306*, 2026.

Shuwei Yue and Minchen Wei. Effective cross-sensor color constancy using a dual-mapping strategy. *Journal of the Optical Society of America A*, 41(2):329–337, 2024.

A Complete Experimental Protocol and Data Splits

A.1 Data Splits

We split the synthetic benchmark along three dimensions: reflectance, illuminant, and RGB camera. HASCID contains 765 reflectance scenes of size $512 \times 512 \times 204$, sampled at 204 nonuniform wavelengths from 397.32 to 1003.58 nm. We divide the scenes into 612/76/77 for training/validation/test; no reflectance scene appears in more than one split.

After preprocessing, we retain 1,608 measured illuminant SPDs from two datasets: 1,506 real-light SPDs from Royer (Royer, 2023) and 102 measured SPDs from Barnard et al. (Barnard et al., 2002). Together they cover diverse natural and artificial illumination, including outdoor daylight, LED, fluorescent, incandescent, and daylight-simulating sources. All spectra cover 380–780 nm; the Royer and Barnard spectra are sampled at 1 nm and 4 nm, respectively. We split the 1,608 SPD records into 1,286/161/161 for training/validation/test. For each reflectance scene, we sample 20 illuminants without replacement from the corresponding split, giving 12,240/1,520/1,540 synthesized images.

Table 4: Dataset statistics and the combination rule for each split.

Split	Reflectances	Illuminants	Per scene	Images	RGB-camera pool
Train	612	1286	20	12,240	36 source cameras
Validation	76	161	20	1,520	14 held-out cameras
Test	77	161	20	1,540	14 held-out cameras

A.2 Camera Split

We use a manufacturer-held-out protocol. The 36 source cameras comprise 22 Canon cameras, three Fujifilm cameras, and 11 Nikon cameras:

- **Canon (22):** Digital Rebel XTi, EOS-1D X Mark II, EOS-1Ds Mark III, EOS-1Ds Mark II, EOS 100D, EOS 200D, EOS 200D II, EOS 5DS, EOS 5D, EOS 5D Mark III, EOS 5D Mark II, EOS 5D Mark IV, EOS 600D, EOS M, EOS R10, EOS R5, EOS R5m2, EOS R6, EOS R6m2, EOS RP, EOS R, and PowerShot S90.
- **Fujifilm (3):** GFX 100, X-T3, and X-T4.
- **Nikon (11):** D200, D3300, D5100, D5300, D7000, D700, D70, D800E, D810, D850, and Z f.

The 14 held-out target cameras comprise 11 Sony cameras, one Panasonic camera, and two Hasselblad cameras:

- **Sony (11):** DSC-RX100M4, ILCE-6400, ILCE-7CM2, ILCE-7M3, ILCE-7M4, ILCE-7RM2, ILCE-7RM3, ILCE-7RM4, ILCE-7SM2, ILCE-7SM3, and ILCE-9.
- **Panasonic (1):** DC-GX9.
- **Hasselblad (2):** L1D-20c and L2D-20c.

Validation and test use the same 14 target-camera identities but different reflectance scenes and illuminant directories. No target camera is used to train the network.

A.3 Target-Camera Calibration Set

Stage 1 uses 60 MSI–RAW pairs per target camera. We shuffle the HASCID training reflectances and select 60 distinct scenes. Each scene is paired with one randomly selected SPD from the training illuminant pool and one random Zurich homography. Before sampling, we discard extreme

illuminants whose crest factor is at least 4. All 14 target cameras share the same 60 reflectance–illuminant–homography triples and the same MSI observations. Consequently, the calibration set contains $60 \times 14 = 840$ pairs.

A.4 Spectral Sampling and Integration

Before rendering, illuminant SPDs, RGB sensitivities, multispectral sensitivities, and CIE color-matching functions are aligned to the 204-point HASCID grid. Downsampling uses band averaging between adjacent target-band boundaries to preserve narrow spectral peaks; empty bands fall back to linear interpolation. Upsampling uses linear interpolation, and values outside the source wavelength range are zero.

All sensor responses use trapezoidal integration with the true wavelength spacing. For wavelengths $\{\lambda_k\}_{k=1}^{L}$, the endpoint weights are $w_1 = (\lambda_2 - \lambda_1)/2$ and $w_L = (\lambda_L - \lambda_{L-1})/2$, while an interior weight is $w_k = (\lambda_{k+1} - \lambda_{k-1})/2$. A channel response is

$$I_{p,c} = \sum_{k=1}^{L} w_k R_p(\lambda_k) E(\lambda_k) S_c(\lambda_k). \tag{13}$$

The CSS and white-point working grid is 400–720 nm at 5 nm intervals (65 samples). The SPD supervision target is 400–720 nm at 10 nm intervals (33 samples) and is ℓ_2-normalized. The original RGB sensitivities cover 380–780 nm at 5 nm intervals. The multispectral sensitivity table covers 353–1049 nm and has 16 channels centered at 400, 415, 435, 465, 500, 530, 555, 580, 630, 650, 680, 705, 745, 765, 835, and 855 nm.

A.5 Color-Correction Evaluation Protocol

XYZ-to-Lab conversion and pixelwise CIEDE2000. All methods are evaluated in linear D65 CIE XYZ; the metric does not include XYZ-to-sRGB conversion, gamma encoding, tone mapping, or display-gamut clipping. For both the prediction and target, we use the fixed $Y = 1$ D65 Lab reference white

$$\mathbf{X}_n = (X_n, Y_n, Z_n) = (0.950456,\ 1.0,\ 1.089058). \tag{14}$$

For a pixel $\mathbf{X} = (X, Y, Z)$, let $(x_r, y_r, z_r) = (X/X_n, Y/Y_n, Z/Z_n)$ and

$$f(u) = \begin{cases} u^{1/3}, & u > (6/29)^3, \\ u/[3(6/29)^2] + 4/29, & \text{otherwise.} \end{cases} \tag{15}$$

We then compute

$$L^* = 116 f(y_r) - 16, \quad a^* = 500[f(x_r) - f(y_r)], \quad b^* = 200[f(y_r) - f(z_r)]. \tag{16}$$

The pixelwise error is the standard CIEDE2000 distance (Sharma et al., 2005), with $k_L = k_C = k_H = 1$.

Normalization and clipping. During synthesis, RAW is divided by the largest channel of the camera response to a unit-reflectance surface under the scene illuminant. The D65 XYZ target is independently divided by the largest tristimulus component of a unit-reflectance surface under D65. RAW and target XYZ are clipped to $[0, 1]$ after rendering and again after homography warping and cubic resizing. In contrast, predicted XYZ is not clipped before conversion to Lab: negative values enter the linear branch of Eq. 15, and values above one are retained. No sRGB out-of-gamut clipping is involved. All predictions and targets entering the reported statistics are finite.

CCT and dual-illuminant DNG interpolation. An RGB baseline supplies only a camera-space white point, so CCT and the coefficient in Eq. 3 are recovered iteratively from that white point and the camera's two ColorMatrices. In our convention, t is the weight of the Standard-A endpoint. Starting from $t = 0.5$, each iteration computes

$$\mathrm{CM}(t) = t\,\mathrm{CM}_{\mathrm{A}} + (1 - t)\,\mathrm{CM}_{\mathrm{D65}}, \qquad \mathbf{q} = \mathrm{CM}(t)^{-1}\mathbf{w}, \tag{17}$$

followed by

$$(x, y) = \left(\frac{q_X}{q_X + q_Y + q_Z}, \frac{q_Y}{q_X + q_Y + q_Z} \right). \tag{18}$$

We estimate CCT with the McCamy approximation

$$n = \frac{x - 0.3320}{0.1858 - y}, \qquad T = 449n^3 + 3525n^2 + 6823.3n + 5520.33, \tag{19}$$

and update the reciprocal-temperature interpolation weight as

$$t = \mathrm{clip}_{[0,1]} \left(\frac{T^{-1} - T_{\mathrm{D65}}^{-1}}{T_{\mathrm{A}}^{-1} - T_{\mathrm{D65}}^{-1}} \right), \quad T_{\mathrm{A}} = 2856\ \mathrm{K}, \quad T_{\mathrm{D65}} = 6504\ \mathrm{K}. \tag{20}$$

We run at most 30 iterations and stop when the absolute change in t is below 10^{-6}. The same coefficient is then used for elementwise interpolation,

$$\mathrm{FM}(t) = t\,\mathrm{FM}_{\mathrm{A}} + (1 - t)\,\mathrm{FM}_{\mathrm{D65}}. \tag{21}$$

The ForwardMatrix output is in the D50 profile connection space. We therefore apply a Bradford adaptation from the interpolated ForwardMatrix white to D65. This is a DNG-style implementation of the reciprocal-CCT dual-illuminant interpolation rule.

B Identifiability and Stability of CSS Calibration

Response-domain identifiability. For a scene-mean radiance $\bar{\mathbf{w}}$, the MSI observation and pseudoinverse estimate are

$$\bar{\mathbf{m}} = \mathbf{S}_{\mathrm{MS}}\bar{\mathbf{w}}, \qquad \tilde{\mathbf{w}} = \mathbf{S}_{\mathrm{MS}}^{+}\bar{\mathbf{m}} = \mathbf{S}_{\mathrm{MS}}^{+}\mathbf{S}_{\mathrm{MS}}\bar{\mathbf{w}}. \tag{22}$$

Here $\mathbf{P} = \mathbf{S}_{\mathrm{MS}}^{+}\mathbf{S}_{\mathrm{MS}}$ is the orthogonal projector onto the MSI row space. Since $\mathbf{S}_{\mathrm{MS}} \in \mathbb{R}^{16\times 204}$ has rank 16 and a 188-dimensional null space, recovering the physical radiance is underdetermined and non-unique; $\tilde{\mathbf{w}} = \mathbf{P}\bar{\mathbf{w}}$ is instead the minimum-norm, MSI-consistent projection. Stage 1, however, requires response equivalence rather than full radiance recovery. For the sensitivity $\mathbf{s}_c$ of an RGB channel, let $\mathbf{I} \in \mathbb{R}^{204\times 204}$ denote the identity matrix. The induced response error is

$$\mathbf{s}_c^{\mathsf{T}}\bar{\mathbf{w}} - \mathbf{s}_c^{\mathsf{T}}\tilde{\mathbf{w}} = \mathbf{s}_c^{\mathsf{T}}(\mathbf{I} - \mathbf{P})\bar{\mathbf{w}}. \tag{23}$$

Here, $(\mathbf{I} - \mathbf{P})\bar{\mathbf{w}}$ is the radiance component discarded by the MSI projection, and Eq. 23 isolates its contribution to the RGB-response error. Stage 1 does not treat $\tilde{\mathbf{w}}$ as a physical radiance estimate; it uses $\hat{\mathbf{S}}\tilde{\mathbf{w}}$ only to predict the mean RGB response for fitting the observed RAW mean. Its relevant criterion is therefore response preservation rather than spectral reconstruction. We evaluate both below.

Validation of the radiance projection. Table 5 evaluates Eq. 22 on the 60 calibration-scene mean radiances. For any reference vector $\mathbf{x}$ and estimate $\tilde{\mathbf{x}}$, relative error is $\|\tilde{\mathbf{x}} - \mathbf{x}\|_2/\|\mathbf{x}\|_2$, and angular error is the angle between the two vectors. Radiance metrics compare $\tilde{\mathbf{w}}$ with $\bar{\mathbf{w}}$; response metrics apply the MSI or RGB sensitivity operator to both radiances using the same trapezoidal integration weights as the renderer.

The radiance-level results confirm that the pseudoinverse is not reconstructing the physical spectrum: the relative ℓ_2 error is 6.875% on average and reaches 33.087% in the worst scene, while the mean spectral angle is 3.956°. By construction, however, $\mathbf{S}_{\mathrm{MS}}\mathbf{P} = \mathbf{S}_{\mathrm{MS}}$, so the MSI response is preserved up to numerical precision (1.37×10^{-15} mean relative error). More importantly, the spectral discrepancy has little effect on the target RGB cameras: the mean response angle is only 0.0438°, with a 90th percentile of 0.0678° and a maximum of 0.1346°. The mean relative RGB-response error is likewise only 0.116% and remains below 0.262% for every scene. The small relative error additionally shows that response magnitude is well preserved. These results justify using $\tilde{\mathbf{w}}$ as a response-equivalent intermediate representation, rather than interpreting it as a recovered physical radiance.

Stability of CSS estimation. The analysis above concerns the radiance used to form the response equations; estimating CSS from a finite set of these equations can still be ill-conditioned. We stabilize the estimate by restricting each 65-dimensional CSS correction to the k-dimensional PCA subspace in Eq. 4, centering it at the target-specific ColorMatrix anchor, using spectrally diverse calibration scenes, and imposing second-order spectral smoothness.

Table 5: Error induced by projecting the true scene-mean radiance through $\mathbf{S}_{\mathrm{MS}}^{+}\mathbf{S}_{\mathrm{MS}}$.

Quantity	Mean	Median	90th percentile	Maximum
Radiance relative ℓ_2 error	6.875%	4.551%	12.498%	33.087%
Radiance spectral angle (°)	3.956	2.608	7.181	19.321
MSI-response relative ℓ_2 error	1.37×10^{-15}	1.34×10^{-15}	1.63×10^{-15}	1.98×10^{-15}
RGB-response angle (°)	0.0438	0.0414	0.0678	0.1346
RGB-response relative ℓ_2 error	0.116%	0.113%	0.210%	0.262%

Table 6: Stage 1 stability with respect to PCA dimension, smoothness weight, and response-domain iterations. Each sweep varies one factor while fixing the others to $k = 10$, $\lambda_{\mathrm{sm}} = 3 \times 10^{-4}$, and 30 iterations. Entries are angular errors in degrees; lower is better.

PCA dimension			Smoothness			Iterations		
k	CSS angle	Imaging error	λ_{sm}	CSS angle	Imaging error	Iter.	CSS angle	Imaging error
3	5.128	0.335	0	32.105	5.311	0	6.937	2.216
5	**3.844**	0.185	10^{-6}	8.309	0.358	1	5.596	0.243
8	4.395	0.161	10^{-5}	6.276	0.251	2	5.239	0.210
10	4.338	**0.157**	10^{-4}	4.483	0.166	3	5.038	0.195
15	4.605	0.159	3×10^{-4}	**4.338**	0.157	5	4.798	0.182
			10^{-3}	4.576	**0.154**	10	4.539	0.169
			3×10^{-3}	4.973	0.175	20	4.382	0.160
			10^{-2}	5.606	0.246	30	4.338	0.157
						50	**4.323**	**0.155**

Table 6 summarizes the stability analysis. Moderate PCA dimensions are sufficient: $k = 5$ gives the lowest CSS spectral error ($3.844°$), whereas the default $k = 10$ gives the lowest imaging error ($0.157°$). Smoothness is essential: removing it increases the CSS error from $4.338°$ to $32.105°$. The response error decreases from $2.216°$ at the ColorMatrix anchor to $0.157°$ after 30 iterations and $0.155°$ after 50 iterations; we therefore use 30 iterations as the accuracy–computation trade-off.

C Number of Target-Camera Calibration Pairs

We vary the number of MSI–RAW calibration pairs per target camera as $N_{\mathrm{cal}} \in \{0, 1, 5, 10, 20, 40, 60\}$. At $N_{\mathrm{cal}} = 0$, CSS is estimated only from the target camera's DNG ColorMatrices (ColorMatrix-only). All Stage 2 and Stage 3 weights and the test set are fixed; only the recovered CSS is changed. For $N_{\mathrm{cal}} \in \{1, 5, 10, 20, 40\}$, we report the mean and standard deviation over five random calibration subsets. The 0- and 60-pair settings are deterministic.

With only 1–5 pairs, calibration remains sensitive to scene selection. The imaging error decreases, although the CSS spectral angle can temporarily increase; with one pair, this instability can also degrade the downstream color error. This shows that curve-shape and image-formation errors are not equivalent. With 10 pairs, white-point error and ΔE_{00} reach $1.389°$ and 2.456; with 20 pairs, they reach $1.362°$ and 2.385, close to the 60-pair results of $1.349°$ and 2.376.

D Additional Results for Illuminant Spectrum Estimation

Tab. 8 presents the complete experimental results for illuminant spectrum estimation, including comparisons with statistical, reconstruction-based, and learning-based methods. Our method outperforms all compared methods and achieves the best results across all five error metrics. Unlike the Spectrum baseline, which collapses the spectral axis early and uses a fully connected readout over flattened features, our method preserves wavelength-specific features and adopts a lightweight spec-

Table 7: Effect of the number of target-camera calibration pairs. Lower is better for all metrics.

N_{cal}	CSS angle (°)	Imaging error (°)	White-point error (°)	ΔE_{00}
0 (ColorMatrix-only)	6.937	2.216	2.659	4.915
1	12.377 ± 0.173	1.764 ± 0.505	2.257 ± 0.408	5.457 ± 1.129
5	8.364 ± 0.876	0.682 ± 0.224	1.572 ± 0.172	3.115 ± 0.546
10	5.750 ± 0.399	0.284 ± 0.048	1.389 ± 0.016	2.456 ± 0.044
20	4.825 ± 0.084	0.184 ± 0.012	1.362 ± 0.005	2.385 ± 0.008
40	4.559 ± 0.116	0.163 ± 0.005	1.356 ± 0.004	2.377 ± 0.007
60	**4.338**	**0.157**	**1.349**	**2.376**

Table 8: Illuminant spectrum estimation, evaluated by the spectral-angle error between the predicted and ground-truth SPDs. Best results are shown in bold and second-best results are underlined.

Method	**SPD angular error↓**				
	Mean	**Med.**	**Tri.**	**B25%**	**W25%**
White-Patch + FFNN (Erba et al., 2024a)	14.55	13.33	13.13	6.06	25.99
Gray-World + FFNN (Erba et al., 2024a)	14.29	12.91	12.83	5.86	25.79
Gray-Edge (1st) + FFNN (Erba et al., 2024a)	14.64	13.14	13.09	6.18	26.20
PWIR (Robles-Kelly & Wei, 2018)	8.06	6.31	6.78	3.00	16.09
Spectrum (Oh et al., 2026)	<u>5.28</u>	<u>4.35</u>	<u>4.48</u>	<u>2.19</u>	<u>10.11</u>
Ours	**4.71**	**3.60**	**3.79**	**1.51**	**10.03**

tral readout. This design reduces the mean/median angular errors from $5.28°/4.35°$ to $4.71°/3.60°$, while reducing the parameter count from 16.3M to 0.68M.

E Virtual-Chart Construction and Sensitivity Analysis

Protocol. The fixed virtual chart is drawn exclusively from the HASCID training split. Using fixed seed 1, we sample 100 training scenes and 50 pixels per scene without replacement, giving the default $N_R = 5{,}000$. Spectra are resampled to the 400–720 nm, 5-nm grid. Because the HASCID scene splits are disjoint (Sec. A), no reflectance scene used by the virtual chart occurs in validation or test. For the size analysis, we repeat this procedure for five seeds, globally shuffle each 5,000-spectrum bank, and use nested prefixes of size $N_R \in \{24, 100, 500, 1000, 5000\}$.

Number of virtual reflectances. Table 9 reports mean test-image ΔE_{00}, averaged over five chart seeds; $\pm$ denotes the standard deviation across seeds. The complete model is stable from $N_R = 500$: its mean changes by less than 0.002 for $N_R \in \{500, 1000, 5000\}$, showing that the result is not caused by a particular virtual-chart selection.

CCM and reflectance-library controls. Table 10 compares the conventional DNG baseline, a directly learned CCM, and our physical-plus-residual model with different reflectance libraries. The directly learned CCM uses the same RAW, CSS, and white-point inputs as our residual branch, but predicts the complete matrix without a physical initialization. The directly learned CCM reaches a mean ΔE_{00} of 2.565, while introducing the physical initialization and residual refinement reduces it to 2.395. This improvement holds across all 14 held-out cameras, indicating that the physical anchor provides a more transferable basis than unconstrained direct CCM regression. Using the 24 standard ColorChecker reflectances gives 2.487, close to the 2.395 obtained with our larger virtual chart. This result suggests that a compact, readily available standard reflectance chart can provide a practical alternative for deployment when a large reflectance library is unavailable.

Table 9: Sensitivity to the number of HASCID-training reflectances. "Full" denotes the physical anchor plus the residual trained with upstream-error injection. Results are averaged over five chart seeds.

N_R	Physical CCM $\Delta E_{00} \downarrow$	Full $\Delta E_{00} \downarrow$
24	2.639 ± 0.064	2.441 ± 0.027
100	2.630 ± 0.011	2.405 ± 0.010
500	2.622 ± 0.023	2.394 ± 0.004
1000	2.614 ± 0.026	2.394 ± 0.005
5000	2.593 ± 0.011	2.392 ± 0.008

Table 10: CCM and reflectance-library controls.

Method	Reflectance library	N_R	Mean ΔE_{00}	Median ΔE_{00}
DNG interpolated	–	–	3.85	3.62
Directly learned CCM	–	–	2.57	2.11
Physical + residual + error injection	ColorChecker-24	24	2.49	2.05
Physical + residual + error injection	HASCID train	5,000	**2.40**	**1.92**

F Oracle Analysis of Physical-Input Errors

Our framework forms a cascade in which the recovered camera spectral sensitivity (CSS) conditions Stage 2, and the resulting illuminant spectral power distribution (SPD) and camera-space white point are subsequently used to construct the physical CCM. We conduct two complementary experiments to trace how errors propagate through this cascade. First, we replace the Stage 1 CSS with its ground truth and re-run Stage 2 to measure the resulting changes in SPD and white-balance estimation. Second, we perform controlled Oracle replacements when constructing the physical CCM and quantify how much color-restoration error each replacement recovers. We restrict this Oracle analysis to the deterministic physical CCM, since replacing individual inputs with ground truth would create mixed input distributions for the learned residual; its practical contribution under the normal predicted-input distribution is evaluated separately in Table 3.

Effect of CSS accuracy on Stage 2. To quantify how Stage 1 CSS error propagates into Stage 2, we run the frozen Stage 2 estimator twice on the same test images: once with the predicted CSS and once with the corresponding ground-truth CSS. All image inputs, model weights, and evaluation settings are otherwise unchanged. As shown in Table 11, replacing the predicted CSS with its ground truth reduces the mean SPD error by only 0.004° (0.08%) and the mean AWB error by 0.031° (2.31%). The small changes indicate that the recovered CSS already provides a sufficiently accurate camera representation for Stage 2, and that Stage 1 error contributes only weakly to the remaining SPD and white-point errors.

Oracle analysis of the physical CCM. We next evaluate how errors in CSS, SPD, and white-point estimation affect the closed-form color mapping. Starting from the full predicted configuration, we replace one quantity at a time with its ground truth and recompute both the virtual chart and $\mathbf{M}_{\text{phys}}$. For the CSS Oracle, we also re-run Stage 2 with the ground-truth CSS and use its updated SPD and white-point predictions, as in Table 11. The all-GT setting replaces all three quantities simultaneously. Every configuration uses the same test images, image resolution, reflectance library, model weights, and spectral sampling.

As shown in Table 12, the cascaded CSS intervention only reduces the mean ΔE_{00} from 2.586 to 2.550, consistent with the small changes in Stage 2 reported in Table 11. In contrast, replacing the predicted SPD with its ground truth reduces the error to 0.834, while replacing all physical inputs with ground truth further reduces it to 0.716. These results show that illuminant-spectrum estimation

Table 11: Effect of replacing the predicted CSS with ground truth in Stage 2. Results are reported as mean (median) values over the same 1,540 test samples and 14 held-out target cameras; lower is better.

CSS input	SPD angular error (°) ↓	White-point angular error (°) ↓
Predicted CSS	4.71 (3.59)	1.35 (0.95)
Ground-truth CSS	4.71 (3.58)	1.32 (0.94)

Table 12: Oracle analysis of physical-input errors in the closed-form dynamic CCM. "Pred." and "GT" denote predicted and ground-truth quantities, and "Re-pred." indicates that Stage 2 is re-run with the ground-truth CSS. Results are mean (median) per-image CIEDE2000 over the same 1,540 test samples; lower is better.

Setting	CSS	SPD	WB	ΔE_{00} ↓
Full predicted inputs	Pred.	Pred.	Pred.	2.59 (2.11)
CSS cascade Oracle	GT	Re-pred.	Re-pred.	2.55 (2.08)
SPD Oracle	Pred.	GT	Pred.	0.83 (0.71)
WB Oracle	Pred.	Pred.	GT	2.59 (2.11)
All-GT inputs	GT	GT	GT	0.72 (0.62)

is the dominant source of error in the current physics-based color mapping, whereas CSS recovery and the closed-form CCM construction contribute much less.

The WB Oracle produces the same physical-CCM error as the full predicted configuration. This is because the white-balance gains are used consistently to white-balance the RAW image and to recompute the dynamic CCM, allowing the CCM to absorb the corresponding channel scaling. Thus, WB does not act as an independent error source in the purely linear physical path, although it remains relevant to learned modules conditioned on the white-balanced representation.

G Evaluation on Real-World Data

We evaluate the complete pipeline on the Beyond-RGB dataset (Glatt et al., 2024) using its official training, validation, and test splits. The dataset provides paired 16-channel multispectral (MS) images, smartphone RAW images, measured illuminant spectra, and ColorChecker observations under 36-band spectral sampling (380–730 nm at 10 nm). Because the dataset contains only two smartphone cameras, we evaluate our pipeline in a single-camera setting using the Samsung subset. Beyond-RGB does not provide smartphone CSS ground truth; its MS camera response is a 16×36 matrix. These properties require a dataset-specific Stage 1 calibration, while Stages 2–3 retain the proposed inference procedure.

G.1 Stage 1: CSS Calibration under Real MS Observations

The released HDF5 MS data have undergone black-level subtraction, saturation normalization, bad-pixel interpolation, lens-shading correction, and bilateral denoising. Consequently, they need not exactly follow the ideal linear forward model used for synthetic training. We re-extract the 24 chart patches using the official 4×4 MS mosaic order and shrink each polygon to 60% of its original extent. After fitting one exposure scalar per scene, the measured HDF5 responses still differ from responses synthesized from the measured reflectance, illuminant, and MS sensitivity by 22.5% in relative error and 4.99° in spectral angle. The nonzero singular values of the supplied MS response matrix have a condition number of 275.9, further amplifying this model discrepancy.

We use imaging error, defined as the angular error between the response obtained by integrating the measured SPD with the estimated CSS and the measured RAW response of the ColorChecker neutral-gray patch.

We compare two Stage 1 inputs using the same thirty calibration scenes, CSS anchor, validation protocol, and measured Samsung RAW responses. In the matched-model control, ColorChecker reflectance and the measured illuminant first form the ideal patch radiance. We then synthesize an ideal MSI using the released MS CSS and recover the radiance with its pseudoinverse before applying Stage 1 refinement. In the direct-radiance setting, the ColorChecker reflectance–illuminant product is used directly as the radiance. Both settings use real Samsung RAW responses for CSS refinement and are evaluated on held-out real scenes. Their similar mean imaging errors, 1.111° and 0.980°, show that matched MSI formation and pseudoinversion introduce only a small loss, supporting the validity of our calibration method on real data.

Table 13: Stage 1 calibration on Beyond-RGB. Lower is better.

Stage 1 calibration signal	Imaging error (°)		
	Mean	Median	P95
Color-matrix anchor	1.600	1.510	2.391
Ideal MSI → MS-CSS pseudoinverse	1.111	0.962	1.841
Reflectance × measured SPD (direct radiance)	**0.980**	**0.757**	**1.799**

G.2 Stage 2: Chart-Free Illuminant Estimation

At inference, Stage 2 receives only the chart-free 16-channel MS image and the corresponding Samsung RAW image; the white-target image and chart are not used. The network predicts a scale-normalized 36-dimensional SPD and white-balance representation. Training uses the measured SPD and the white-target neutral-gray response as supervision. Table 14 compares the direct white-balance head with white points obtained by integrating either the predicted or measured SPD through the Stage 1 CSS. With CSS fixed, the gap between predicted and measured SPD quantifies the effect of illuminant estimation on the recovered white point. The learned white-balance head is slightly more accurate than direct integration of the predicted SPD.

Table 14: Stage 2 results on Beyond-RGB. Lower is better.

Predicted quantity	Compared with	Angular error (°)		
		Mean	Median	P95
SPD	Measured SPD	3.668	2.278	7.869
White-balance head	Measured neutral-gray RAW	1.405	0.935	2.786
Predicted SPD integrated with Stage 1 CSS	Measured neutral-gray RAW	1.509	1.075	2.941
Measured SPD integrated with Stage 1 CSS	Measured neutral-gray RAW	0.935	0.754	1.767

G.3 Stage 3: Dynamic Color Restoration

Stage 3 follows the scene-adaptive CCM construction in Section 3.4: the recovered CSS and predicted SPD are used to build a physical CCM from fixed reference reflectances, without using the chart observed in the test image. Because Beyond-RGB does not provide dense XYZ ground truth for the phone images, we evaluate only the physical CCM and do not apply the learned residual refinement; the ColorChecker reference values are used only for evaluation. We evaluate 157 scenes after excluding two saturated captures. Samsung chart patches are re-extracted with 60% polygon shrinkage and compared with the official post-2014 X-Rite ColorChecker targets, chromatically adapted from D50 to D65 using Bradford adaptation. We average CIEDE2000 over 24 patches within each scene and then over scenes.

Table 15: Stage 3 real-data color restoration on 157 Beyond-RGB scenes. The metric is the mean per-scene CIEDE2000 over the standard 24 patches.

Method	$\Delta E_{00} \downarrow$
DNG interpolation + predicted SPD/AWB	5.442
Dynamic physical CCM + predicted SPD/AWB	4.674
Dynamic physical CCM + measured SPD/AWB	4.048
Per-scene chart-fitted 3×3 matrix	**3.655**

The proposed dynamic transform improves over DNG interpolation by 0.768 ΔE_{00}. Replacing the predicted SPD with the measured SPD lowers the error to 4.048, giving the physical-pipeline upper bound under the recovered CSS. The empirical oracle directly fits a separate 3×3 matrix from each test scene's observed 24 RGB patches to their reference XYZ values and evaluates those same patches; it is therefore non-deployable and represents a more favorable, chart-supervised bound. Its residual error of 3.655 reflects the limited expressiveness of a linear 3×3 mapping together with acquisition noise, spatial illumination variation, patch-extraction error, and mismatch between the generic reference values and the physical chart.

H Additional Qualitative Results

We provide additional qualitative comparisons that isolate camera variation and cover both the single-camera and cross-camera evaluation settings. The value overlaid on each restored image is its mean pixelwise ΔE_{00}; lower is better.

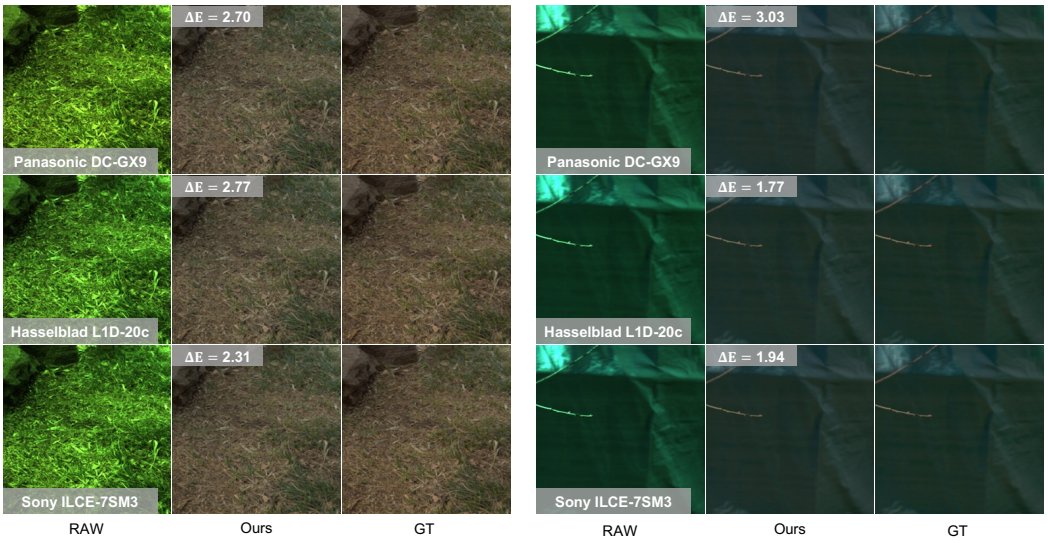

Figure 5: Controlled cross-camera results under fixed scene content and illumination. Within each example, only the target RGB camera changes across rows, while the scene, light source, and multispectral observation remain fixed. Although the RAW appearance varies substantially across the Panasonic DC-GX9, Hasselblad L1D-20c, and Sony ILCE-7SM3, our restored results remain consistent and closely match the camera-independent ground truth.

Figure 5 directly visualizes the role of the MSI as a camera-agnostic spectral bridge. Changing the RGB sensor alters the RAW response because each camera has a different spectral sensitivity, even though the underlying scene radiance is unchanged. After incorporating the recovered camera sensitivity, our framework maps these different sensor responses to similar colors. The small variation in ΔE_{00} across cameras further indicates that the restoration is not tied to one camera's RAW color space.

Figure 6 contains challenging scenes dominated by large regions with similar colors, including foliage, fabric, and ground surfaces. Such regions provide limited within-image chromatic diversity and can make illumination and surface color difficult to disentangle from RGB alone. The multispectral observation supplies denser spectral evidence, allowing our method to remove the strong green cast while preserving the intended surface colors. In these examples, it also remains competitive with MSCC despite MSCC being trained specifically for the test camera, whereas our model transfers without target-camera retraining.

As shown in Figure 7, the RGB baselines leave color shifts that vary with the target sensor. Our method produces colors closer to the ground truth across all three held-out cameras, reducing the displayed ΔE_{00} to 0.48, 2.06, and 1.50, respectively. Together with the controlled comparison in Figure 5, these results demonstrate that the proposed physical factorization handles both difficult scene content and camera-dependent RAW responses.



Figure 6: Additional single-camera comparisons on the Sony ILCE-7RM2. From left to right: RAW input, C5, the camera-specific MSCC model, our result, and ground truth. Our method accurately restores colors in spatially extensive, nearly uniform-color regions.

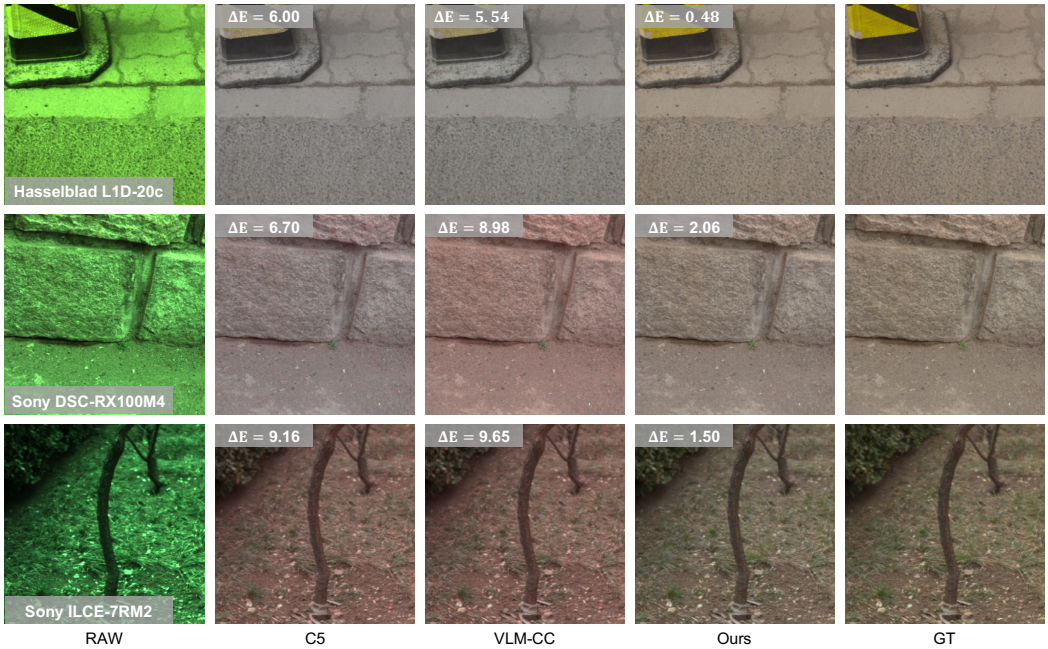

Figure 7: Additional cross-camera comparisons on three held-out cameras. From left to right: RAW input, C5, VLM-CC, our result, and ground truth. The scenes are captured with the Hasselblad L1D-20c, Sony DSC-RX100M4, and Sony ILCE-7RM2, none of which is used to train our framework.

I PER-CAMERA RESULTS

We complement the result in Table 1 with the complete per-camera breakdown in Table 16. Here, N denotes the number of held-out test images for each camera. Despite the variation in CSS recovery, downstream performance is comparatively stable: AWB error spans $1.109°$–$1.597°$ and ΔE_{00} spans 2.123–2.560. Notably, the Hasselblad L1D-20c has the largest CSS error but still achieves low AWB and color errors, suggesting that the subsequent stages tolerate camera-dependent CSS estimation errors across the evaluated manufacturers.

Table 16: Per-camera results on all 14 held-out cameras.

Manufacturer	Camera	N	CSS ($^\circ$)	SPD ($^\circ$)	AWB ($^\circ$)	ΔE_{00}
Sony	DSC-RX100M4	99	5.01	4.73	1.29	2.39
Sony	ILCE-6400	105	4.46	4.51	1.24	2.12
Sony	ILCE-7CM2	111	3.36	4.99	1.50	2.56
Sony	ILCE-7M3	113	3.19	5.13	1.34	2.36
Sony	ILCE-7M4	126	3.43	4.45	1.44	2.37
Sony	ILCE-7RM2	113	4.00	5.23	1.60	2.56
Sony	ILCE-7RM3	95	4.13	4.95	1.58	2.47
Sony	ILCE-7RM4	105	3.54	4.97	1.53	2.41
Sony	ILCE-7SM2	109	4.35	4.43	1.18	2.20
Sony	ILCE-7SM3	122	3.48	4.63	1.31	2.37
Sony	ILCE-9	112	3.95	4.85	1.38	2.52
Panasonic	DC-GX9	106	5.69	4.19	1.21	2.33
Hasselblad	L1D-20c	100	6.97	4.68	1.11	2.22
Hasselblad	L2D-20c	124	5.17	4.30	1.19	2.40

J Cross-Camera Consistency of the White-Balance Spectral Representation

The white-balance head outputs a nonnegative, ℓ_2-normalized representation $\hat{\mathbf{E}}_{\mathrm{wb}} \in \mathbb{R}^{65}$. Its supervision is applied only after projection through the recovered CSS:

$$\hat{\mathbf{l}} = \hat{\mathbf{S}} \cdot \hat{\mathbf{E}}_{\mathrm{wb}}. \tag{24}$$

The representation $\hat{\mathbf{E}}_{\mathrm{wb}}$ does not receive direct SPD supervision. We therefore evaluate both its cross-camera consistency and its discrepancy from the measured physical illuminant spectrum.

Counterfactual cross-camera protocol. We randomly sample 30 test scenes and use all 14 held-out cameras from Sony, Panasonic, and Hasselblad. For each scene, we keep the reflectance, measured illuminant, MSI observation, and geometric homography fixed, and re-render the RGB RAW observation using the ground-truth CSS of every camera. This produces $30 \times 14 = 420$ scene–camera observations in which only the target RGB camera changes. We run the same trained checkpoint using either the ground-truth CSS or the estimated CSS used at deployment. For each scene, we compute the spectral angle between every unordered pair of camera-conditioned WB representations $\hat{\mathbf{E}}_{\mathrm{wb}}$. We average these angular errors over all $30\binom{14}{2} = 2{,}730$ same-scene camera pairs; a smaller value means that changing the RGB camera produces less change in the output representation. We additionally report two complementary metrics: SPD angular error measures the angle from $\hat{\mathbf{E}}_{\mathrm{wb}}$ to the measured SPD, and white-point angular error measures downstream AWB accuracy after CSS projection.

Table 17: Counterfactual cross-camera analysis on 30 test scenes and all 14 held-out cameras. Results are mean (median) angles in degrees; lower is better.

CSS input	Same-scene pairwise feature angle (°) ↓	SPD angular error (°) ↓	White-point angular error (°) ↓
Ground-truth CSS	0.485 (0.272)	30.445 (27.774)	1.338 (0.812)
Predicted CSS	0.590 (0.348)	30.444 (27.780)	1.382 (0.871)

Interpretation. The small $0.485°$ angle with ground-truth CSS shows that changing the camera leaves the predicted WB feature almost unchanged. The model therefore maps camera-dependent RAW observations of the same scene to a shared representation, which is the key behavior required for cross-camera white-balance transfer.

With predicted CSS, the mean angle increases only slightly to $0.590°$, showing that this cross-camera consistency is retained under the actual deployment setting. Although the representation is far from the measured physical SPD ($30.444°$), projecting it through the target-camera CSS recovers the camera white point with a mean error of $1.382°$. These results indicate that $\hat{\mathbf{E}}_{\mathrm{wb}}$ is not a physical illuminant estimate, but a camera-invariant feature specialized for transferring white balance across unseen cameras.