ABC-Former: Auxiliary Bimodal Cross-domain Transformer with Interactive Channel Attention for White Balance

Anonymous CVPR submission

Paper ID 3154

Abstract

The primary goal of white balance (WB) for sRGB images is to correct inaccurate color temperatures, making images exhibit natural, neutral colors. While existing WB methods achieve reasonable results, they are limited by the global color adjustments applied during a camera's post-sRGB processing and the restricted color diversity in current datasets. This often leads to suboptimal color correction, particularly in images with pronounced color shifts. To address these limitations, we propose an Auxiliary Bimodal Cross-domain Transformer (ABC-Former) that enhances WB correction by leveraging complementary knowledge from multiple modalities. ABC-Former employs two auxiliary models to extract global color information from CIELab and RGB color histograms, complementing the primary model's sRGB input processing. We introduce an Interactive Channel Attention (ICA) module to facilitate cross-modality knowledge transfer, integrating calibrated color features into image features for more precise WB results. Experimental evaluations on benchmark WB datasets show that ABC-Former achieves superior performance, outperforming state-of-the-art WB methods.

1. Introduction

White balance (WB) correction aims to make cameras consistently reproduce stable and accurate colors under varying color temperatures and lighting conditions. However, the camera's Image Signal Processing (ISP) often introduces color casts in final sRGB images due to inaccurate or personalized WB settings applied to the input Raw-sRGB images. These color casts may potentially reduce the accuracy of downstream tasks like image classification and semantic segmentation, where precise color representation is essential for reliable performance. As a result, WB correction to remove such color casts has drawn much attention in recent research.

Much work has been devoted to achieving effective WB inside the camera's processing unit. Within the Image Signal Processing (ISP) pipeline, Raw-WB techniques were proposed to deal with color shifts in raw images caused by lighting conditions in the scene [11, 18]. These approaches [11, 18] estimate the scene's illuminant color to adjust the raw image for color correction. However, due to the ISP's non-linear rendering processes, Raw-WB methods often struggle to fully address color shifts in the final sRGB output [1].

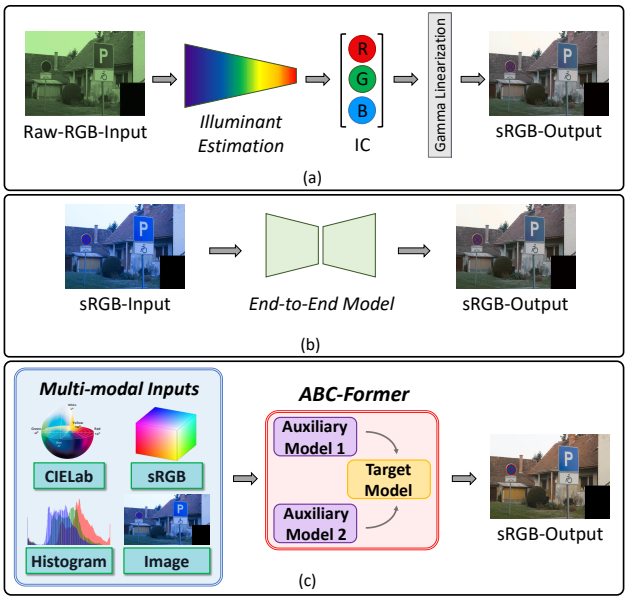

Figure 1. (a) Raw-WB methods first predict the illuminant color of the scene and then adjust the raw-RGB image color through gamma linearization using this predicted illuminant. (b) DNN-based sRGB-WB methods directly process the input sRGB image with an end-to-end model to obtain the WB result. (c) Our proposed ABC-Former converts the input image into multiple modalities, which serve as inputs for both auxiliary and primary models to improve the accuracy of WB correction.

Several sRGB White Balance (sRGB-WB) methods have been proposed to address such color shifts stemming from imprecise WB setups set in the camera's ISP, broadly cat-

egorized into exemplar-based and Deep Neural Network (DNN)-based methods [2, 14, 15]. KNN [3] and Mixed-WB [4] are exemplar-based methods, which classify a limited set of images from the Rendered WB dataset [3] and match the most suitable nonlinear mappings for WB correction. DNN-based methods like DEEP-WB [2] process the single-illuminant input image with a convolutional neural network to correct its color temperature. WBFlow [14] employs a convolutional reversible flow architecture to extract pseudo-raw features, enhancing the quality of feature extraction for sRGB color correction. Another DNN-based method, SWBNet [15], utilizes a transformer-based architecture and operates in the DCT domain to identify and adjust color temperature-sensitive features, aiming to improve accuracy in WB correction. Although these methods have shown promising results, they rely mainly on color and scene information from input images, missing the opportunity to leverage a broader range of data representations.

Our work leverages alternative representations from different modalities, such as color histograms, to learn global color temperature and achieve effective white balance correction. While images provide spatial and color information in a visual format, histogram representations capture the distribution of color intensities across channels without spatial details. Since per-channel color histograms do not fully preserve color relationships between pixels, we go beyond sRGB-input images by exploring both sRGB and chromaticity histograms, allowing us to extract more implicit global color information from limited training data and enhancing the learning of color characteristics across various modalities. Inspired by [21], we propose an Auxiliary Bimodal Cross-domain Transformer architecture (ABC-Former), which includes two auxiliary models to extract corrected global color information from both CIELab and sRGB color histograms, along with a target model to process the sRGB input image. To facilitate target WB correction through the auxiliary models, we introduce an Interactive Channel Attention module (ICA) that uses re-parameterization techniques to transfer modality-complementary knowledge from calibrated color information to the target sRGB model, allowing it to adaptively reweight image features for more accurate WB results. Our contributions are listed as follows:

- We propose ABC-Former, which utilizes modality-complementary knowledge from global color information in histogram-based features via auxiliary models to enhance the accuracy of WB correction in the primary sRGB model.
- To enable effective cross-modality knowledge transfer to the primary model, we introduce the ICA module to reweight sRGB image features, leading to improved WB results.
- Experimental results on public benchmark WB datasets demonstrate that the proposed ABC-Former performs favorably against State-Of-The-Art (SOTA) sRGB-WB methods.

2. Related Works

Raw-WB Approaches. The WB module in a camera's ISP is designed to render raw images with accurate color temperature settings, aiming to eliminate undesirable color shifts caused by varying lighting conditions. Traditional WB methods estimate the global color of the light source and apply a uniform gain coefficient to convert raw images to final sRGB images. However, these methods [6, 7, 11, 16, 19] assume consistent color temperature across the scene, making them ineffective under mixed lighting conditions. Furthermore, because these methods alter the original sRGB image, they are unsuitable for further adjusting the resulting sRGB output.

sRGB-WB Approaches. To address the limitations of traditional WB methods, recent research has shifted towards extending color correction beyond the ISP imaging stage with sRGB-WB approaches to produce final white balance (WB) results. sRGB-WB methods can be broadly categorized into exemplar-based methods [3, 4] and DNN-based methods. Exemplar-based methods rely on trained nonlinear mappings to correct color casts in images. For instance, in [3], images with similar color distributions are identified using histogram features from the training set, and their corresponding correction matrices are combined to create a non-linear color correction matrix that maps the input image's colors to those of the target. Mixed-WB [4] proposes rendering the input image with various predefined white balance settings to produce weighting maps corresponding to those color temperatures. These maps are averaged to correct the image, aligning it with the ideal white balance. However, these methods rely heavily on predefined training data, which limits their adaptability; without a diverse range of scenes and lighting conditions in the training data, these methods often struggle to generalize effectively to unseen images.

DNN-based methods like WBFlow [14] utilize neural flow to guarantee reversibility, allowing for the lossless transformation of color-cast sRGB images back into a pseudo-raw feature space for linear white balancing. Additionally, WBFlow integrates a camera transformation module within this space, enabling adaptation to various cameras through few-shot learning and enhancing generalization. SWBNet [15] suppresses temperature-sensitive low-frequency information and applies a color temperature contrast loss to align features of the same scene across different temperatures, reducing white balance instability. It employs adaptive weights to correct multiple color temperature shifts within an image, achieving accurate WB under mixed lighting conditions. While effective, these meth-

ods rely solely on information within the image modality. **Multimodal Training Approaches.** Unimodal training involves training a model using data from a single modality. Typically, the model is trained and tested on the same or a similar task within a shared domain. While effective within its scope, unimodal training often struggles to generalize across different modalities, limiting its adaptability in diverse scenarios. In contrast, multimodal training methods enable models to learn simultaneously from multiple data modalities. This multimodal data can be either strongly correlated (paired data) or weakly correlated (unrelated data). An example of strongly correlated data is CLIP [17], which uses contrastive learning to map features extracted by image and text encoders, requiring additional paired data from multiple modalities. The Multimodal Pathway Transformer (M2PT) [21] is an approach that trains multiple models with weakly correlated data, using re-parameterization techniques to allow the target model to inherit the capabilities of a pre-trained auxiliary model, thereby capturing information across different modalities. While this approach enables the integration of weakly correlated data, it also raises the cost of collecting diverse data modalities and the pre-training requirements for auxiliary models.

Inspired by M2PT, we opt for multimodal training to capture underlying global color information from sRGB training images in different modalities for enhancing WB correction performance. Unlike M2PT's using strongly uncorrelated data, we train sRGB images to learn corrected histogram-based global color features from their color and chromaticity histograms. The comparison among conventional DNN-based Raw-WB and sRGB-WB methods and the proposed ABC-Former is shown in Figure 1.

3. Proposed Method

The proposed ABC-Former contains three transformer models: two auxiliary transformers that learn to correct color and chromaticity histograms and one primary transformer as a target model that takes the input sRGB image and outputs the final WB result by leveraging cross-modality global color knowledge transferred by the auxiliary models. Since M2PT [21] deals with uncorrelated data from different modalities, it requires an additional tokenization process to convert different data into a unified format to allow the auxiliary and target models to use the same architecture for re-parametrization. In contrast, since our data from different modalities are strongly correlated, we introduce **Interactive Channel Attention (ICA)** module to utilize condensed histogram-based features from different modalities to effectively transfer modality-complementary knowledge to correct color temperature variations for improved color accuracy. The overall architecture is depicted in Figure 2.

3.1. Auxiliary Model — PDFformer

In most prior sRGB-WB works [2–4, 14], the focus is primarily on images within the sRGB domain and the local pixel modality. These approaches often limit the performance due to insufficient attention to global color temperature, which is crucial for guiding sRGB image features in effective WB correction. Therefore, incorporating global color information from alternative modalities could enhance WB accuracy by providing a broader context for color adjustments across the entire image. Given an input sRGB image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the image's height and width, and the three channels represent the RGB color components, we first transform the input sRGB image $\mathbf{I}$ into the CIELab color space, represented as $\mathbf{I_{Lab}} \in \mathbb{R}^{H\times W\times 3}$, where the three channels correspond to the L^*, a^*, and b^* components. To capture the global color temperature of an sRGB image while keeping model complexity low, we convert $\mathbf{I}$ into its probability density function (PDF) form, denoted as $\mathbf{H_{sRGB}} \in \mathbb{R}^{L\times 3}$, where $L = 256$ represents the number of histogram bins per channel. Likewise, we transform $\mathbf{I_{Lab}}$ into its PDF form, referred to as $\mathbf{H_{Lab}} \in \mathbb{R}^{L\times 3}$, with $L = 256$, providing a histogram-based representation for each color space.

As previously mentioned, the entire framework was devised with the inclusion of two auxiliary models to improve the performance of the target model. Global color PDF data, $\mathbf{H^{sRGB}}$ and $\mathbf{H^{Lab}}$, representing color and chromaticity information from the sRGB and CIELab domains, respectively, serve as distinct inputs to the auxiliary models. Each auxiliary model employs a 1-D transformer architecture, referred to as PDFformer, with a shared architecture but trained separately. Initially, the input histograms, $\mathbf{H^{sRGB}}$ or $\mathbf{H^{Lab}}$, pass through a one-dimensional convolutional layer to produce histogram features $\mathbf{H_0^A} \in \mathbb{R}^{L\times C}$, where $\mathbf{A} \in [\mathbf{sRGB}, \mathbf{Lab}]$ and C represent the feature dimensions. These features $\mathbf{H_0^A}$ are subsequently fed into a transformer-based U-shape structure with PDFformer blocks that facilitate upsampling and downsampling within the encoder and decoder pathways. Each PDFformer block consists of two sequences of Layer Normalization (LN), Channel Attention (CA) [10], and a feed-forward Multilayer Perceptron (MLP) [8], arranged in the following order:

$$\begin{aligned} \hat{\mathbf{H}}_\mathbf{i}^\mathbf{A} &= \mathrm{CA}(\mathrm{LN}(\mathbf{H}_{\mathbf{i-1}}^\mathbf{A})) + \mathbf{H}_{\mathbf{i-1}}^\mathbf{A}; \\ \mathbf{H}_\mathbf{i}^\mathbf{A} &= \mathrm{GELU}(\mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{H}}_\mathbf{i}^\mathbf{A}))) + \hat{\mathbf{H}}_\mathbf{i}^\mathbf{A}, \end{aligned} \tag{1}$$

where GELU$(\cdot)$ denotes the GELU activation function, and $\mathbf{i}$ is the block index, starting from 1. PDFformer has K PDFformer blocks in the encoding path, followed by a bottleneck stage, which is also a PDFformer block. These blocks are interconnected via downsampling, implemented through a 4×1 convolution with a stride of 2 and channel doubling. This process yields an output of $\mathbf{H_{K+1}^A} \in$

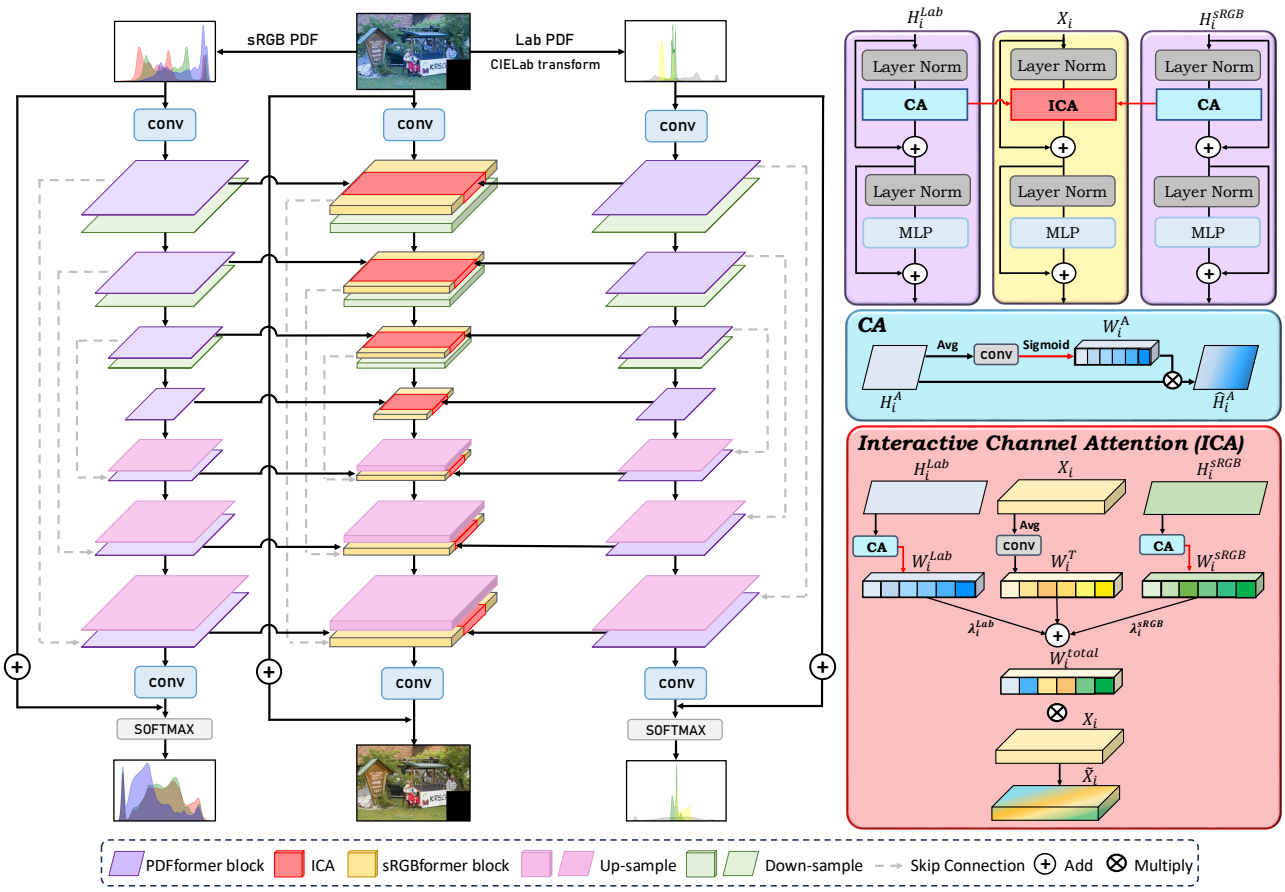

Figure 2. The framework of our ABC-Former consists of two key components: **Auxiliary models (PDFformers)** and **Target model (sRGBformer)**. The auxiliary models utilize histogram information from the sRGB and CIELab color spaces augmented from the input to learn color information across different domains and modalities. The target model's ICA module consolidates the auxiliary models' information into the target model to produce the final WB results.

$\mathbb{R}^{\frac{L}{2^K} \times 2^K C}$. Following the bottleneck, the decoding path contains K PDFformer blocks, with upsampling and channel reduction applied between blocks. The first two blocks halve the number of channels, while the remaining blocks quarter them, as in [20]. Upsampling is achieved using a 2×1 transposed convolution with a stride of 2. Additionally, each block in the decoding path takes the output from the previous block and concatenates it with the corresponding output from the encoding path of the same spatial size. The final output, $\mathbf{H^A} \in \mathbb{R}^{L \times 2C}$, passes through a convolutional layer with a residual connection to the input histograms and a softmax function to produce the corrected color or chromaticity histograms $\mathbf{H_c^A} \in \mathbb{R}^{L \times 3}$ as:

$$\mathbf{H_c^A} = \text{SOFTMAX}(\mathbf{Conv_{3\times1}}(\mathbf{H^A_{2K+1}}) + \mathbf{H^A_0}) \quad (2)$$

3.2. Target model — sRGBformer

To achieve a color-calibrated rendition of the input image, we adopt a vision transformer architecture called sRGBformer to address and correct color deviations. sRGBformer systematically processes the features of the input image $\mathbf{X_0}$ to produce the final WB output image. Like PDFformer, sRGBformer employs a transformer-based U-shaped architecture composed of sRGBformer blocks with upsampling and downsampling in the encoding and decoding paths. A key difference lies in the integration of cross-modality knowledge obtained from the auxiliary models, providing corrected global color information as a guide. To achieve this integration, we replace the CA in each sRGBformer block with our proposed **Interactive Channel Attention (ICA)**, which incorporates knowledge from the auxiliary models through dedicated pathways. Each sRGBformer block consists of two sets of LN, ICA, and an MLP, arranged in the following order:

$$\begin{aligned} \hat{\mathbf{X}}_\mathbf{i} &= \text{ICA}(\text{LN}(\mathbf{X_{i-1}})) + \mathbf{X_{i-1}}; \\ \mathbf{X_i} &= \text{GELU}(\text{MLP}(\text{LN}(\hat{\mathbf{X}}_\mathbf{i}))) + \hat{\mathbf{X}}_\mathbf{i}, \end{aligned} \quad (3)$$

where $\mathbf{X_i}$ is the image features produced by the sRGBformer block at i-th level. Both the encoder and decoder

are composed of K sRGBformer blocks, connected through a bottleneck stage that consists of one sRGBformer block. In the encoder, downsampling and channel doubling are achieved by a 4×4 convolution with a stride of 2. In the decoder, upsampling and channel halving are performed by a 2×2 transposed convolution with a stride of 2. As in PDFformer, each decoder block receives the output from the previous block, concatenated with the corresponding output from the encoder of the same spatial size. The final decoder output $\mathbf{X_{2K+1}} \in \mathbb{R}^{H \times W \times 2C}$ passes through a 3×3 convolutional layer with a residual connection to the input image, producing the final WB image $\mathbf{X_c} \in \mathbb{R}^{H \times W \times 3}$.

Interactive Channel Attention: The goal of ICA is to enable knowledge transfer from auxiliary models with different modalities to the target sRGBformer via dedicated pathways, facilitating more accurate WB correction. In the target model, sRGBformer, each encoder and decoder block containing an ICA module corresponds to respective blocks in encoders and decoders of the auxiliary models. First, $\mathbf{X_i}$ undergoes average pooling along the channel dimension (Avg(.)), followed by a convolution operation, producing the vector $\mathbf{W_i^T} \in \mathbb{R}^{1\times1\times C}$ at the i-th level of the sRGBformer block. The vectors $\mathbf{W_i^{sRGB}}, \mathbf{W_i^{Lab}} \in \mathbb{R}^{1\times1\times C}$ are obtained by feeding $\mathbf{H_i^{sRGB}}$ and $\mathbf{H_i^{Lab}}$ into the CA module within their corresponding PDFformer's block, which include the average pooling operation Avg(.), convolution, the unsqueeze operation, and the excitation operation (sigmoid activation) at the i-th level. Next, using the cross-modal re-parameterization technique [21], we introduce learnable parameters $\lambda_\mathbf{i}^\mathbf{sRGB}$ and λ_i^{Lab} for $\mathbf{W_i^{sRGB}}$ and $\mathbf{W_i^{Lab}}$, respectively, and combine them with $\mathbf{W_i^T}$ along the channel dimension to obtain $\mathbf{W_{total}}$. At last, $\mathbf{X_i}$ is multiplied channel-wise by the scalar weighting vector $\mathbf{W_i^{total}}$ to produce the re-weighted sRGB features $\tilde{\mathbf{X}}_\mathbf{i}$, as follows:

$$\begin{aligned} \mathbf{W_i^T} &= \mathbf{Conv}(\mathrm{Avg}(\mathbf{X_i})); \\ \mathbf{W_i^{total}} &= \mathbf{W_i^T} + \lambda_\mathbf{i}^\mathbf{Lab}\mathbf{W_i^{Lab}} + \lambda_\mathbf{i}^\mathbf{sRGB}\mathbf{W_i^{sRGB}}; \\ \tilde{\mathbf{X}}_\mathbf{i} &= \mathbb{F}_{scale}(\mathbf{W_i^{total}}, \mathbf{X_i}), \end{aligned} \quad (4)$$

where $\mathbb{F}_{scale}(\cdot,\cdot)$ denotes the channel-wise multiplication function, applying scalar weights to the corresponding feature maps. Through ICA, we leverage calibrated global color information extracted from squeezed modality-specific histogram-based features to transfer modality-complementary knowledge, effectively guiding sRGB image features for improved WB correction.

3.3.

Loss Function The proposed framework optimizes two auxiliary models and one target model: two auxiliary PDFformers, which take histogram inputs in a PDF format from the sRGB or CIELab domains, respectively, and the target sRGBformer. The cooperative interaction among the auxiliary and target models is essential for achieving precise WB in the output sRGB image. To train auxiliary models, we employ L2 loss to measure the difference between the PDFs of the color channels and the ground truth as:

$$\begin{aligned} \mathcal{L}_{\text{pdf}}^{sRGB} &= \|\mathbf{H_c^R} - \mathbf{H_{gt}^R}\|_2 + \|\mathbf{H_c^G} - \mathbf{H_{gt}^G}\|_2 + \|\mathbf{H_c^B} - \mathbf{H_{gt}^B}\|_2; \\ \mathcal{L}_{\text{pdf}}^{Lab} &= \|\mathbf{H_c^L} - \mathbf{H_{gt}^L}\|_2 + \|\mathbf{H_c^a} - \mathbf{H_{gt}^a}\|_2 + \|\mathbf{H_c^b} - \mathbf{H_{gt}^b}\|_2, \end{aligned} \quad (5)$$

where $\mathbf{H_c^{sRGB}}$=$[\mathbf{H_c^R}; \mathbf{H_c^G}; \mathbf{H_c^B}]$ and $\mathbf{H_c^{Lab}}$=$[\mathbf{H_c^L}; \mathbf{H_c^a}; \mathbf{H_c^b}]$ denote the corrected RGB and Lab histograms, respectively, as estimated by PDFformers. $\mathbf{H_{gt}^{sRGB}}$ and $\mathbf{H_{gt}^{Lab}}$ refer to the ground-truth histograms. We use L2 loss as it penalizes large deviations more heavily, encouraging the model to align histogram bins across the distribution evenly. Thus, it is preferable to KL divergence, which may overemphasize small-probability bins or become unstable when encountering zero probabilities. For sRGBformer, we apply L1 loss to train it as:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{X_c} - \mathbf{X_{gt}}\|_1, \quad (6)$$

where $\mathbf{X_c}$ is the estimated WB image produced by sRGBformer, and $\mathbf{X_{gt}}$ is the ground truth. The total loss is then defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pdf}}^{sRGB} + \mathcal{L}_{\text{pdf}}^{Lab} + \mathcal{L}_{\text{rec}}$.

4. Experimental Results

Datasets. The commonly used public dataset, the Rendered WB dataset Set1 [3], is divided into three non-overlapping folds. For training, we randomly selected 12,000 rendered sRGB images with different WB settings from two of the folds. For testing, we use the third fold of Set1, also known as Set1-Test (21,046 images), as well as other datasets that have no overlap in scenes or cameras with the training data: Set2 of the Rendered WB dataset (2,881 images) [3] and the Rendered Cube+ dataset (10,242 images) [5]. These are used to evaluate the WB performance of our proposed ABC-Former.

Evaluation Metrics. We employed three common evaluation metrics to assess our results: Mean Square Error (MSE), Mean Angular Error (MAE), and ΔE 2000 [9], which calculate the differences between the predicted white-balanced images and the ground truth. For each metric, we reported the mean, first quantile (Q1), median (Q2), and upper quantile (Q3) of the error. Lower values in these metrics indicate better WB correction performance, consistent with those used in recent works [2–4, 14, 15].

Implementation details. We implemented ABC-Former using PyTorch. During training, we optimize both the auxiliary and target models simultaneously over 350 epochs using the Adam optimizer [13] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for each model. The learning rate is set to 2×10^{-4}, and the embedding feature dimension to 16. During training, we randomly cropped four 128×128 patches from training images as input. In addition, we apply geometric transforma-

Table 1. The quantitative results of our ABC-Former and competing WB methods are evaluated on three public benchmark datasets [3, 5]. The best results are highlighted in red, while the second-best results are highlighted in blue.

Method	MSE ↓				MAE ↓				ΔE 2000 ↓				Size
	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3	MB
Rendered WB Dataset: Set1-Test (21,046 images) [3]													
KNN [3]	**77.49**	13.74	39.62	94.01	3.06°	1.74°	2.54°	3.76°	3.58	2.07	3.09	4.55	21.8
Deep-WB [2]	82.55	13.19	42.77	102.09	3.12°	1.88°	2.70°	3.84°	3.77	2.16	3.30	4.86	16.7
Mixed-WB [4]	142.25	26.81	67.17	164.66	4.07°	2.64°	3.68°	5.16°	4.55	3.00	4.15	5.63	5.1
WBFlow [14]	78.89	**12.99**	**35.09**	**79.35**	**2.67°**	**1.73°**	**2.39°**	**3.24°**	**3.13**	**1.92**	**2.79**	**3.94**	30.2
SWBNet* [15]	111.62	20.61	60.68	137.91	4.11°	2.56°	3.75°	5.22°	4.54	2.73	4.16	5.86	258.8
ABC-Former	**20.47**	**4.65**	**10.02**	**21.05**	**1.99°**	**1.25°**	**1.73°**	**2.33°**	**2.18**	**1.38**	**1.86**	**2.59**	20.2
Rendered WB Dataset: Set2 (2,881 images) [3]													
KNN [3]	171.09	37.04	87.04	190.88	4.48°	2.26°	3.64°	5.95°	5.60	3.43	4.90	7.06	21.8
Deep-WB [2]	124.07	**30.13**	76.32	154.44	3.75°	2.02°	3.08°	4.72°	4.90	**3.13**	4.35	6.08	16.7
Mixed-WB [4]	188.76	48.64	112.32	219.91	4.92°	2.69°	4.10°	6.37°	6.05	3.45	4.92	7.20	5.1
WBFlow [14]	**117.60**	31.25	**61.68**	**143.90**	**3.51°**	**1.93°**	**2.92°**	**4.47°**	**4.64**	3.16	**4.07**	**5.56**	30.2
SWBNet* [15]	219.02	55.45	113.98	236.25	5.46°	3.45°	4.78°	6.63°	6.51	4.39	5.84	8.08	258.8
ABC-Former	**104.31**	**25.55**	**58.61**	**132.90**	**3.39°**	**1.87°**	**2.73°**	**4.30°**	**4.56**	**2.97**	**4.13**	**5.63**	20.2
Rendered Cube+ Dataset (10,242 images) [3, 5]													
KNN [3]	194.98	27.43	57.08	118.21	4.12°	1.96°	3.17°	5.04°	5.68	3.22	4.61	6.70	21.8
Deep-WB [2]	80.46	15.43	33.88	74.42	3.45°	1.87°	2.82°	4.26°	4.59	2.68	3.81	5.53	16.7
Mixed-WB [4]	161.80	16.96	**19.33**	90.81	4.05°	**1.40°**	**2.12°**	4.88°	4.89	**2.16**	**3.10**	6.78	5.1
WBFlow [14]	75.39	**14.22**	30.90	**72.91**	3.34°	1.87°	2.82°	**4.11°**	**4.28**	2.68	3.77	5.21	30.2
SWBNet [15]	**74.35**	20.46	40.04	86.95	**3.15°**	**1.33°**	**2.09°**	4.12°	**4.28**	2.40	3.56	**5.09**	258.8
ABC-Former	**60.60**	**12.15**	**26.92**	**57.20**	**2.99°**	1.63°	2.45°	**3.69°**	**3.95**	**2.35**	**3.40**	**4.86**	20.2

tions, including rotation and flipping, to augment the data.

Quantitative Experimental Results. The quantitative results in Table 1 show that ABC-Former performs favorably against five SOTA methods [2–4, 14, 15] on three public benchmark datasets [3, 5]. These compared methods were evaluated using their publicly available pre-trained models, or results were directly cited from their publications. However, SWBNet [15] was retrained to be tested on the Set1-Test and Set2 datasets, as its code and original results for these datasets were not provided (marked with $*$ in Table 1). SWBNet's scores on the Cube+ Dataset were taken from the original paper. On the Rendered WB Dataset Set1-Test and Set2, ABC-Former achieved the best performance on MSE, MAE, and ΔE 2000, indicating that it effectively removes color casts from images and achieves superior WB correction. On the Rendered Cube+ dataset, ABC-Former also delivered outstanding results, with the lowest mean scores in MSE, MAE, and ΔE 2000. Additionally, ABC-Former maintains a competitive model size, only larger than Deep-WB [2] and Mixed-WB [4]. These results demonstrate ABC-Former's ability to achieve efficient WB correction across various datasets without significantly increasing model complexity, showcasing its robustness and generalization capabilities.

Qualitative Experimental Results. We present the qualitative comparison results on the Rendered WB and Rendered Cube+ datasets in Figure 3 and Figure 4. Observing the color correction results from each method, we see that while KNN [3], Deep-WB [2], Mixed-WB [4], and WBFlow [14] generally reduce color casts in the input images, some color inconsistencies remain across different areas. For example, in Figure 3, these methods tend to correct the colors of objects while neglecting the color accuracy of the sky, resulting in a yellow tint. Similarly, in Figure 4, severe color casts lead to less satisfactory WB results, leaving an overall blue or yellow tone. In contrast, our method, guided by corrected global color information from multiple modalities, achieves a more balanced color correction across the image, producing natural and harmonious results.

Additionally, Figure 5 compares the accuracy of global color distributions in WB results across different methods.

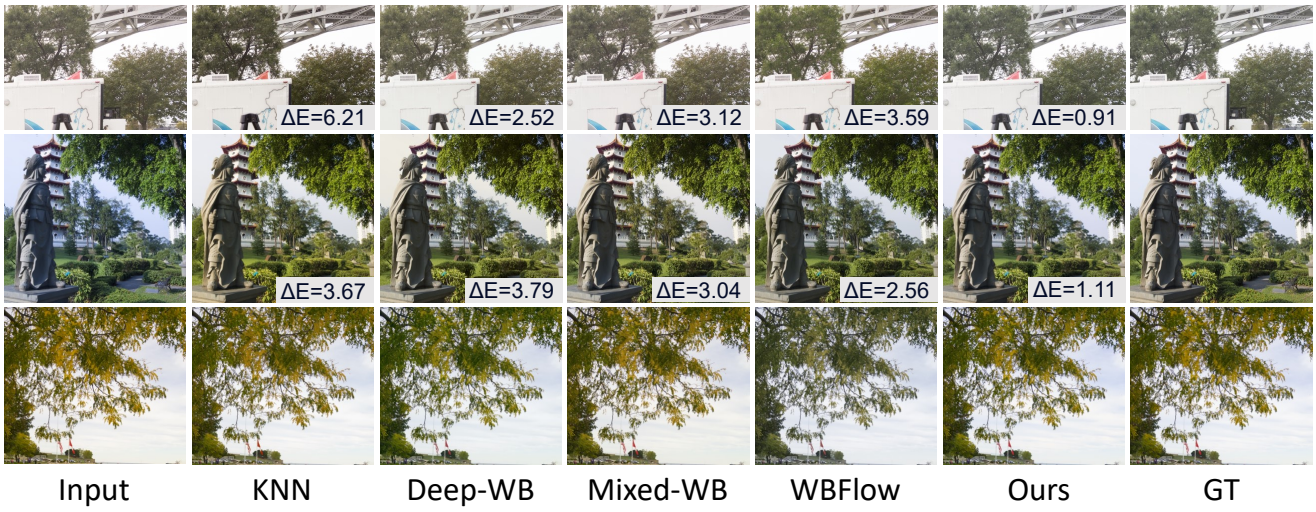

Figure 3. Qualitative comparisons with other sRGB-WB methods on the Rendered WB dataset [3], with the ΔE 2000 indicated in the bottom-right corner of each image.

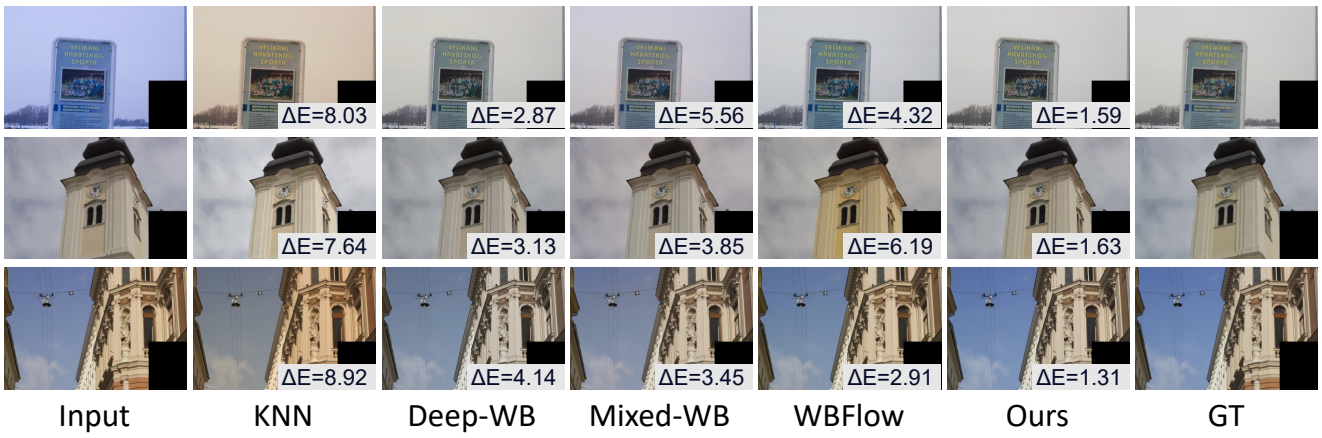

Figure 4. Qualitative comparisons with other sRGB-WB methods on the Rendered Cube+ dataset [5], with the ΔE 2000 displayed in the bottom-right corner of each image.

We use the Bhattacharyya coefficient [12] on red, green, and blue histograms to assess similarity to ground-truth WB images across three benchmark datasets. This coefficient ranges from 0 to 1, with higher values indicating closer alignment with ground truth color distributions. As shown, ABC-Former achieves a higher similarity. The visualization at the bottom of the figure further illustrates an example of the color distributions in WB results obtained by the compared methods, where ABC-Former produces color distributions that are more consistent with the ground truth.

Ablation Studies. In the ablation studies, we analyzed the impact of different combinations of considered modalities used for WB correction on the Rendered Cube+ dataset. We report the mean values across three evaluation metrics, along with model sizes for each combination of auxiliary models. Table 2 shows that using only the target model (sRGBformer) without additional guidance of global color information from other modalities results in limited WB accuracy. Adding a single auxiliary model (e.g., PDF_{sRGB}

Table 2. Ablation studies of ABC-Former w/ and w/o guidance from different modalities on the Rendered Cube+ dataset [5]. Here, sRGB denotes sRGBformer, while PDF_{sRGB}, and PDF_{Lab} represent auxiliary PDFformer models for sRGB and CIELab histograms, respectively.

Modalities	MSE ↓	MAE ↓	ΔE 2000 ↓	Size(MB)
sRGB	81.48	3.70°	4.64	9.8
PDF_{Lab} + sRGB	75.38	3.36°	4.33	15.0
PDF_{sRGB} + sRGB	69.81	3.19°	4.14	15.0
PDF_{sRGB} + PDF_{Lab} + sRGB	**60.60**	**2.99°**	**3.95**	20.2

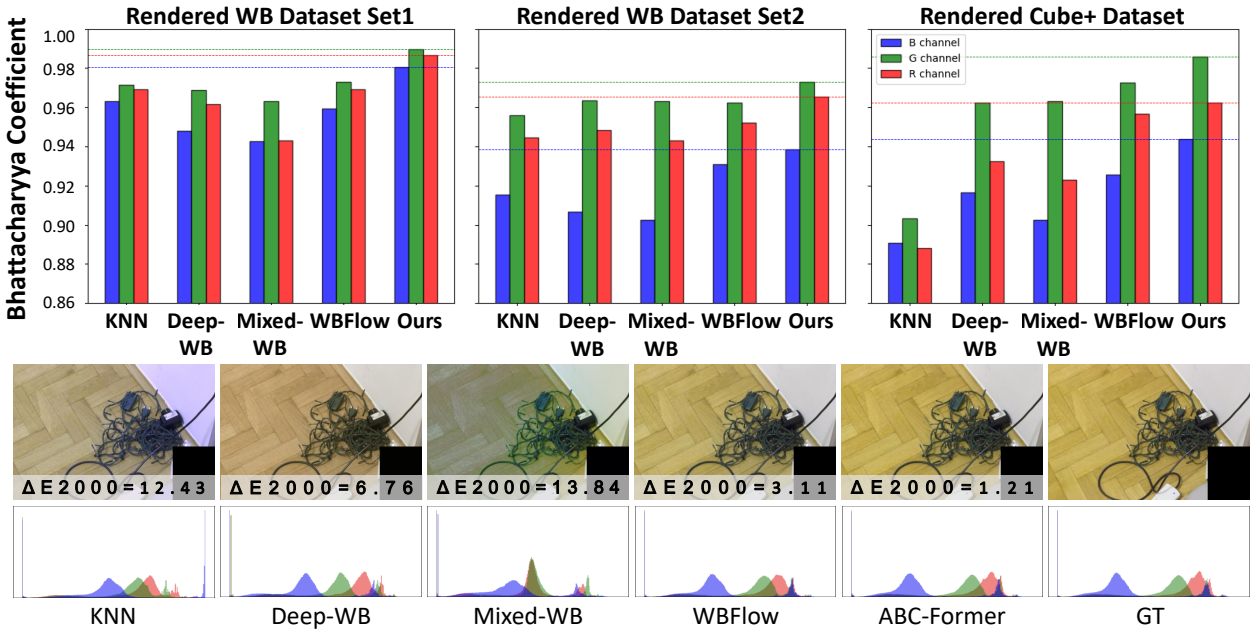

Figure 5. Evaluation of the Bhattacharyya coefficient [12] for color histograms across three benchmark datasets [3, 5], with visualization of global color accuracy for red, green, and blue histograms in existing sRGB-WB methods. ABC-Former leverages global color temperature features from multiple modalities and domains, resulting in more consistent global color distributions throughout the WB process compared to other methods.

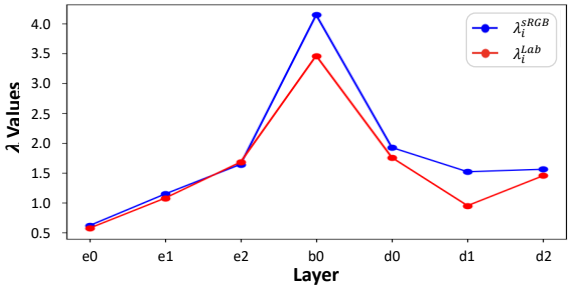

Figure 6. Analysis on learned weights on Rendered Cube+ dataset [5], $\lambda_i^{\mathbf{sRGB}}$ and $\lambda_i^{\mathbf{Lab}}$, for Cross-modality Knowledge Transfer. Here, $i \in \{e_0, e_1, e_2, b_0, d_0, d_1, d_2\}$ represents the level of the sRGBformer block, where e, b, and d denote the encoder, bottleneck, and decoder layers, respectively.

+ sRGB or PDF_{Lab} + sRGB) improves WB performance over the target model alone. Our full ABC-Former design leverages global color temperature features from multiple modalities to guide color adjustment, achieving the best WB accuracy.

Analysis on Learned Weights for Cross-modality Knowledge Transfer. We present the learned weights, $\lambda_i^{\mathbf{sRGB}}$ and $\lambda_i^{\mathbf{Lab}}$, across each level of ABC-Former to demonstrate how modality-complementary knowledge guides WB correction. Figure 6 shows that the influence of calibrated global color information from sRGB and CIELab histogram modalities intensifies toward the bottleneck block of ABC-Former, peaking at the bottleneck. This suggests that high-level WB color histogram-based features from both color and chromaticity modalities, which capture global, semantically rich color information, play a crucial role in reweighting sRGB image features for precise WB correction. Additionally, the sRGB histogram modality has a slightly greater impact on WB correction than CIELab, though the difference is minimal. This indicates that both color modalities contribute collaboratively, facilitating ABC-Former in effectively eliminating color biases.

5. Conclusion

We presented ABC-Former, an Auxiliary Bimodal Cross-domain Transformer designed to advance sRGB WB correction by leveraging complementary information from multiple modalities. ABC-Former employs Interactive Channel Attention to facilitate cross-modality knowledge transfer, allowing the target model to incorporate calibrated color features from both CIELab and RGB histograms. This multimodal approach enables a more nuanced fusion of color information, enhancing the target model's capacity to handle diverse color temperatures and complex scenes with pronounced color shifts. Through extensive experiments, ABC-Former demonstrated superior WB performance, surpassing state-of-the-art methods in both quantitative and qualitative assessments. These results validate ABC-Former's effectiveness in producing more accurate, natural color corrections across varied WB conditions.

References

[1] Mahmoud Afifi and Michael S Brown. What else can fool deep learning? addressing color constancy errors on deep neural network performance. In *ICCV*, pages 243–252, 2019. 1

[2] Mahmoud Afifi and Michael S Brown. Deep white-balance editing. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 1397–1406, 2020. 2, 3, 5, 6

[3] Mahmoud Afifi, Brian Price, Scott Cohen, and Michael S Brown. When color constancy goes wrong: Correcting improperly white-balanced images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1535–1544, 2019. 2, 5, 6, 7, 8

[4] Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. Auto white-balance correction for mixed-illuminant scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1210–1219, 2022. 2, 3, 5, 6

[5] Nikola Banić, Karlo Koščević, and Sven Lončarić. Unsupervised learning for color constancy. *arXiv preprint arXiv:1712.00436*, 2017. 5, 6, 7, 8

[6] Simone Bianco and Claudio Cusano. Quasi-unsupervised color constancy. In *CVPR*, pages 12212–12221. Computer Vision Foundation / IEEE, 2019. 2

[7] Jonathan Cepeda-Negrete and Raul E Sanchez-Yanez. Gray-world assumption on perceptual color spaces. In *Image and Video Technology: 6th Pacific-Rim Symposium, PSIVT 2013, Guanajuato, Mexico, October 28-November 1, 2013. Proceedings 6*, 2014. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[9] Sharma Gaurav. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *COLOR research and application*, 30 (1):21–30, 2005. 5

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 3

[11] Yuanming Hu, Baoyuan Wang, and Stephen Lin. Fc4: Fully convolutional color constancy with confidence-weighted pooling. In *CVPR*, 2017. 1, 2

[12] Thomas Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE transactions on communication technology*, 15(1):52–60, 1967. 7, 8

[13] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[14] Chunxiao Li, Xuejing Kang, and Anlong Ming. Wbflow: Few-shot white balance for srgb images via reversible neural flows. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1026–1034, 2023. 2, 3, 5, 6

[15] Chunxiao Li, Xuejing Kang, Zhifeng Zhang, and Anlong Ming. Swbnet: a stable white balance network for srgb images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1278–1286, 2023. 2, 5, 6

[16] 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 *CVPR*, 2021. 2

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[18] Wu Shi, Chen Change Loy, and Xiaoou Tang. Deep specialized network for illuminant estimation. In *ECCV*, 2016. 1

[19] Joost Van De Weijer, Theo Gevers, and Arjan Gijsenij. Edge-based color constancy. *IEEE TIP*, 2007. 2

[20] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 4

[21] Yiyuan Zhang, Xiaohan Ding, Kaixiong Gong, Yixiao Ge, Ying Shan, and Xiangyu Yue. Multimodal pathway: Improve transformers with irrelevant data from other modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6108–6117, 2024. 2, 3, 5