HVDualformer: Histogram-Vision Dual Transformer for White Balance

Yan-Tsung Peng* **and Guan-Rong Chen**

National Chengchi University
ytpeng@cs.nccu.edu.tw, rogerchen3309@gmail.com

Abstract

Capturing images under different color temperatures can result in color casts, causing the color presented in photos to differ from what is perceived by the human eye. Correcting these color temperature shifts to achieve White Balance (WB) is a challenging task, requiring the identification of variations in color tones from diverse light sources and the removal of color casts. The advent of deep neural networks has significantly advanced the progress of WB methods, evolving from simply identifying the scene illumination color to directly producing a color-corrected image from the color-shifted input. To better map color distributions and scene information from the input to the WB image, we propose **HVDualformer**, an end-to-end histogram-vision dual transformer architecture that can rectify color temperature features from WB color histograms and exploit them to adjust image features to yield accurate WB results. Extensive experimental results on public benchmark datasets demonstrate that the proposed model performs favorably against state-of-the-art methods.

Code — https://github.com/ytpeng-aimlab/HVDualformer

Introduction

Images captured under different lighting conditions often exhibit color variations, known as color casts, which can make them appear bluish, yellowish, or reddish. Unlike the human visual system, which perceives colors consistently, these variations affect visual quality. Correcting color casts to achieve proper White Balance (WB) is crucial for images taken under diverse lighting conditions.

Significant efforts have been made to achieve WB. In a camera's Image Signal Processing (ISP) pipeline, Raw-WB methods are used to eliminate color shifts caused by the scene illuminant in captured raw images (Gijsenij and Gevers 2007; Hu, Wang, and Lin 2017). These methods (Shi, Loy, and Tang 2016; Hu, Wang, and Lin 2017) generally estimate the scene's illuminant color and then apply a 3×3 diagonal Color Correction Matrix (CCM) to the captured raw image based on the estimated illumination for color correction. However, because of the non-linear renderings applied by the ISP, Raw-WB methods often fail to fully correct inconsistent color shifts in the final sRGB image (Afifi and Brown 2019). To address the issue, sRGB-WB methods have been developed to correct color temperature variations caused by inaccurate or personalized WB settings within the camera's ISP (Li et al. 2023). Recently, several sRGB-WB methods have been proposed, which can be categorized into exemplar-based and deep neural network (DNN)-based approaches. Exemplar-based methods, such as KNN (Afifi et al. 2019), compute nonlinear mappings to adjust for varying color temperatures. DNN-based methods, such as DEEP-WB (Afifi and Brown 2020), focus on single-illuminant input and employ a convolutional architecture to map images to the corrected color temperature.

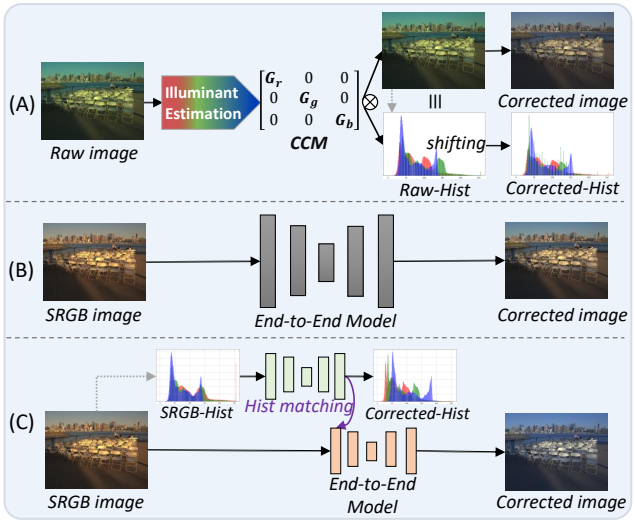

Figure 1: Comparative Overview: (A) Traditional Raw-WB methods correct color shifts by predicting the scene's illuminant and applying a Color Correction Matrix, equivalent to aligning color histograms with WB. (B) End-to-end sRGB-WB correction methods directly process the input sRGB image using a DNN-based model to obtain a WB result. (C) The proposed HVDualformer estimates corrected histogram-based color temperature features and aligns image features based on the corrected histogram features to achieve end-to-end WB correction.

*Corresponding Author.

WBFLOW (Li, Kang, and Ming 2023) employs a convolutional reversible flow architecture to extract pseudo-raw features from the single-illuminant input, enhancing the effectiveness of feature extraction. However, these methods primarily emphasize local pixel adjustments through convolution, often overlooking global color temperature information, which may result in poor overall consistency in color-corrected images. Moreover, there are DNN-based methods (Afifi, Brubaker, and Brown 2022; Kınlı et al. 2023) for multiple-illuminant inputs, aiming to correct images captured under diverse light sources.

This paper proposes HVDualformer, a histogram-vision dual transformer architecture designed to achieve end-to-end WB correction. Traditional WB methods mostly either estimate a single scene illuminant (Shi, Loy, and Tang 2016; Hu, Wang, and Lin 2017) for color correction, which is equivalent to simply shifting color histograms, or they use DNNs for end-to-end WB correction. In contrast, HVDualformer introduces a novel approach that utilizes corrected color histograms to generate global histogram-based color temperature features. These histogram features are used to adjust image features for color correction. Figure 1 demonstrates a comparative overview of traditional WB methods, DNN-based WB methods, and the proposed HVDualformer. HVDualformer comprises two key components: a histogram transformer (**Histoformer**) and a vision transformer (**Visformer**). Specifically, Histoformer learns corrected color temperature features in a histogram-based manner and rectifies image features through Histogram-Specified Feature Transformation (HSFT). Visformer then processes these transformed features to generate the final white-balanced image, effectively removing potential color casts. The proposed HVDualformer not only enhances color accuracy in images but also offers a lightweight solution for correcting color temperature shifts. The main contributions of our work are as follows:

- We introduce a lightweight, streamlined histogram-vision dual transformer architecture that considers corrected histogram-based color temperature features to achieve end-to-end WB correction.
- Our framework is adaptable to handle both single-illuminant and multi-illuminant scenarios (details available on the GitHub page).
- Our HVDualformer achieves state-of-the-art performance on the Rendered WB Dataset Set1 and Set2 (Afifi et al. 2019) and the rendered version of Cube+ Dataset (Banić, Koščević, and Lončarić 2017).

Related Work

Traditional WB Methods. Traditional WB methods can be broadly categorized into two groups: statistic-based and learning-based approaches. Statistic-based methods rely on assumptions or prior conditions derived from observations and analyses of natural scenes and the human visual system. These methods involve calculating the scene's illuminant color to deduce the true colors of objects, as seen in examples like (Cepeda-Negrete and Sanchez-Yanez 2014; Van De Weijer, Gevers, and Gijsenij 2007; Brainard and Wandell 1986). Learning-based methods, on the other hand, involve training models to learn the mapping between input images and their true colors from training data, enabling the output images to closely approximate the true colors, as demonstrated in works, such as (Hu, Wang, and Lin 2017; Lo et al. 2021; Afifi et al. 2019; Bianco and Cusano 2019). While statistic-based methods do not require a large amount of training data, they may struggle in complex scenes and lighting conditions. In contrast, learning-based methods, though requiring substantial paired data for training, can yield better results in scenarios with varying lighting conditions and complex scenes.

sRGB White Balance Methods. As mentioned earlier, traditional WB methods usually rely on shifting image color histograms for WB correction, making it challenging to achieve consistent and accurate results. To address these challenges, recent trends have shifted towards extending the process beyond the ISP imaging stage using DNN-based approaches to generate the final color-corrected image. A primary distinction in DNN-based methods is made between using a single-light source image as input or multiple images with different light sources within the same scene. Single-illuminant methods, such as Deep WB (Afifi and Brown 2020), can achieve decent color correction results. However, it uses a simple U-Net architecture, which primarily extracts scene information, and may reach its limitations in further improving color correction. For multiple-illuminant methods, multiple input images with fixed color temperatures are typically used, including five common WB settings: 5500 K (daylight WB), 2850 K (tungsten WB), 3800 K (fluorescent WB), 6500 K (cloudy WB), and 7500 K (shade WB). Approaches such as Mixed WB (Afifi, Brubaker, and Brown 2022) and Light Style WB (Kınlı et al. 2023) aim to establish correlations between images taken under different color temperatures, effectively addressing scenarios with multiple light sources in a single image to achieve optimal WB results. These methods need input images to adhere to specific criteria, often requiring the use of software tools like Photoshop or the rendering of linear raw DNG image files with different WB settings (e.g., daylight, tungsten, shade) using (Afifi et al. 2019) to meet the input requirements.

Proposed Method

Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, where H and W represent the height and width of the image with three color channels, the proposed HVDualformer aims to correct color temperature shifts and generate the color-corrected output image. HVDualformer consists of two primary components: a histogram transformer (**Histoformer**) and a vision transformer (**Visformer**). Histoformer processes the color histograms of $\mathbf{I}$, denoted as $\mathbf{I}_H$, normalized into a Probability Density Function (PDF) to represent the color distribution. It learns to restore the correct color temperature features for the input image, referred to as histogram features. These histogram features are used to transform the image features by the proposed Histogram-Specified Feature Transformation (HSFT) module, aligning the image features with the corrected histograms.

Next, the corrected image features are concatenated with

the original image features and processed by Visformer to generate a color-balanced representation for the input image. The proposed architecture of HVDualformer is depicted in Figure 2.

Histoformer

As previously mentioned, Histoformer takes the color histogram $\mathbf{I}_H \in \mathbb{R}^{L\times 3}$ of the input image, where $L = 256$ represents the number of bins in the histogram, to learn the corrected color temperature features. To begin with, the input histogram passes through a one-dimensional convolutional layer to generate histogram features $\mathbf{H}_0 \in \mathbb{R}^{L\times C}$, where C is the feature dimension. These embedding features $\mathbf{H}_0$ are then fed into a transformer-based U-shape structure, built with Histoformer blocks that perform upsampling and downsampling in the encoder and decoder paths. Each Histoformer block comprises two sets of Layer Normalization (LN), Channel Attention (CA) (Hu, Shen, and Sun 2018), and feed-forward Multilayer Perceptron (MLP) (Dosovitskiy et al. 2020) in the following sequence:

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

where GELU$(\cdot)$ denotes the GELU activation function, and $\mathbf{i}$ starts from 1. The architecture is depicted in Figure 3(b). Histoformer has K Histoformer blocks in the encoding path, along with a bottleneck stage, which is also a Histoformer block. The blocks are interconnected with downsampling, achieved through a 4×1 convolution with a stride of 2 and channel doubling. This process outputs $\mathbf{H}_{\mathbf{K+1}} \in \mathbb{R}^{\frac{L}{2^K}\times 2^K C}$. After the bottleneck, the decoding path contains K Histoformer blocks and has the upsampling and channel reduction (halving the number of channels for the first two blocks and quartering them for the remaining blocks, as in (Wang et al. 2022)) via 2×1 transposed convolution with a stride of 2 between two blocks. Additionally, each block takes the output from the previous one and concatenates it with the output of the same size from the encoding path. The final output of the decoding path $\mathbf{H}_{\mathbf{2K+1}} \in \mathbb{R}^{L\times 2C}$ passes through two convolutional layers with a residual connection to the input color histogram and a softmax function to produce the corrected color histograms $\mathbf{H}_{\mathbf{c}} \in \mathbb{R}^{L\times 3}$ as:

$$\begin{aligned}\hat{\mathbf{H}}_{\mathbf{c}} &= \mathbf{Conv}_{\mathbf{1\times 1}}(\mathbf{H}_{\mathbf{2K+1}}),\\ \mathbf{H}_{\mathbf{c}} &= \text{SOFTMAX}(\mathbf{Conv}_{\mathbf{3\times 1}}(\hat{\mathbf{H}}_{\mathbf{c}}) + \mathbf{H}_{\mathbf{0}}),\end{aligned} \tag{2}$$

where $\hat{\mathbf{H}}_{\mathbf{c}} \in \mathbb{R}^{L\times C}$ represents the corrected histogram-based color temperature features, which are used to adjust image features for the corrected color distributions in the HSFT module.

Histogram-Specified Feature Transformation (HSFT)

Previously, we utilized Histoformer to generate the corrected histogram-based color temperature features (referred to as histogram features). Next, we introduce the Histogram-Specified Feature Transformation (HSFT) module to adjust the image features for WB correction. First, the input image passes through a convolutional layer to extract image features as $\mathbf{X}_0 = \mathbf{Conv}_{\mathbf{3\times 3}}(\mathbf{I})$, where $\mathbf{X}_0 \in \mathbb{R}^{H\times W\times C}$. The HSFT module then uses the corrected histogram features $\hat{\mathbf{H}}_{\mathbf{c}}$ to refine the input features $\mathbf{X}_0$ through the histogram matching function Hist-Matching (Coltuc, Bolon, and Chassery 2006) as

$$\begin{aligned}\mathbf{H}_{\mathbf{X}} &= \text{PDF}(\text{Norm}(\mathbf{X}_{\mathbf{0}})),\\ \mathbf{X}_{\mathbf{P}} &= \text{Hist-Matching}(\text{SOFTMAX}(\hat{\mathbf{H}}_{\mathbf{c}}), \mathbf{H}_{\mathbf{X}}, \mathbf{X}_{\mathbf{0}})\\ \hat{\mathbf{X}}_{\mathbf{H}} &= \text{De-Norm}(\mathbf{X}_{\mathbf{P}}),\end{aligned} \tag{3}$$

where $\text{Norm}(\mathbf{X}_{\mathbf{0}}) = (L\frac{\mathbf{X}_{\mathbf{0}}}{\max(\mathbf{X}_{\mathbf{0}})})$ with rounding to the nearest integer, and PDF stands for a probability density function that converts the normalized image features into a probability distribution for histogram specification. The Hist-Matching function aligns the histograms of the image features $\mathbf{H}_{\mathbf{X}}$ with the target corrected color histogram $\hat{\mathbf{H}}_{\mathbf{c}}$ by transforming the image features $\mathbf{X}_{\mathbf{0}}$ into $\mathbf{X}_{\mathbf{P}}$. At last, $\mathbf{X}_{\mathbf{P}}$ is de-normalized by dividing L and multiplying by $\max(\mathbf{X}_{\mathbf{0}})$ to obtain the corrected image features $\hat{\mathbf{X}}_{\mathbf{H}}$. The architecture of HSFT is depicted in Figure 3(a).

Visformer

To produce a color-corrected version of the input image, we exploit a vision transformer architecture named Visformer to remove color shifts. Visformer processes the input image features $\mathbf{X}_{\mathbf{0}}$ concatenated with the color-temperature-corrected image features $\hat{\mathbf{X}}_{\mathbf{H}}$ obtained from HSFT to generate the final WB output image.

Specifically, the input to Visformer is $\mathbf{S}_{\mathbf{0}} = \text{Concat}(\mathbf{X}_{\mathbf{0}}, \hat{\mathbf{X}}_{\mathbf{H}})$. Similar to Histoformer, Visformer follows a transformer-based U-shaped structure, backboned using Visformer blocks with upsampling and downsampling in the encoding and decoding paths. A Visformer block consists of two sets of LN, CA, and MLP arranged in the following sequence:

$$\begin{aligned}\hat{\mathbf{S}}_{\mathbf{i}} &= \text{CA}(\text{LN}(\mathbf{S}_{\mathbf{i-1}})) + \mathbf{S}_{\mathbf{i-1}}\\ \mathbf{S}_{\mathbf{i}} &= \text{GELU}(\text{MLP}(\text{LN}(\hat{\mathbf{S}}_{\mathbf{i}}))) + \hat{\mathbf{S}}_{\mathbf{i}}.\end{aligned} \tag{4}$$

Both the encoder and decoder paths contain K Visformer blocks, connected by a bottleneck stage comprising a single Visformer block. In the encoder, downsampling and channel doubling are achieved via a 4×4 convolution with a stride of 2. In the decoder, upsampling and channel halving are performed by 2×2 transposed convolution with a stride of 2. As in Histoformer, the input to each block in the decoder consists of the previous output concatenated with the corresponding output from the encoder of the same size. The final output of the decoder $\mathbf{S}_{\mathbf{2K+1}} \in \mathbb{R}^{H\times W\times 4C}$ passes through a 3×3 convolutional layer with a residual connection to the input image, producing the final WB image $\mathbf{S}_{\mathbf{c}} \in \mathbb{R}^{H\times W\times 3}$.

Loss Functions

The proposed framework involves optimizing two networks: Histoformer and Visformer. The synergistic interaction between these two networks is crucial for achieving accurate

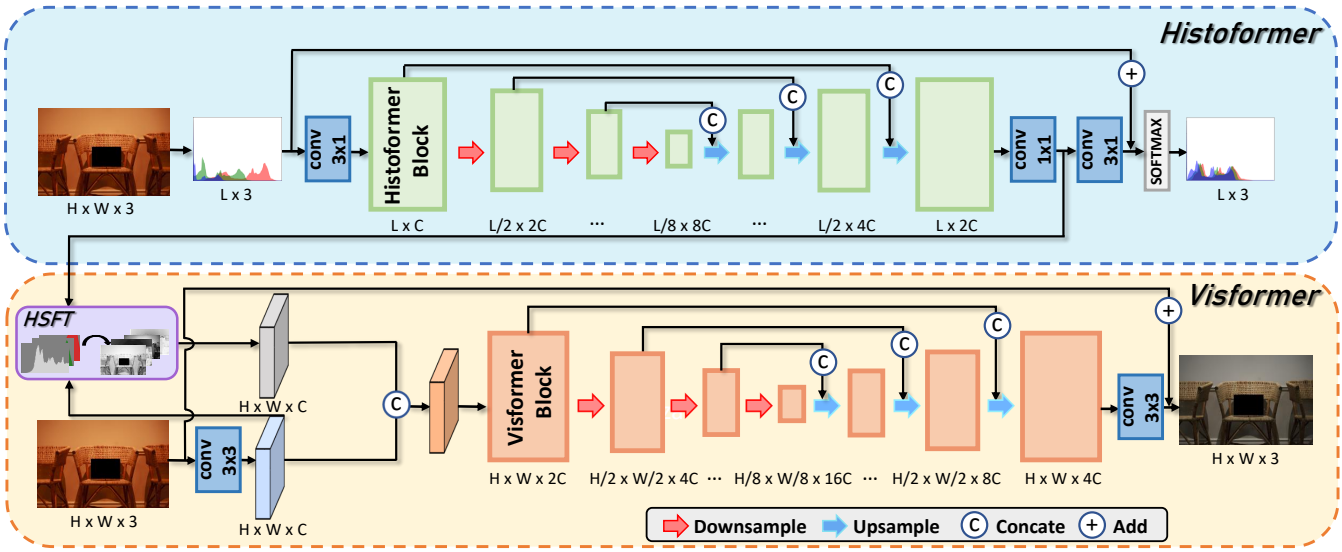

Figure 2: The proposed HVDualformer architecture includes two key components: **Histoformer** and **Visformer**. Histoformer is designed to learn the corrected color temperature features, which are then used to transform the input image features through the Histogram-Specified Feature Transformation (HSFT) module. Visformer processes the corrected image features, concatenated with the original image features, to produce the final WB result.

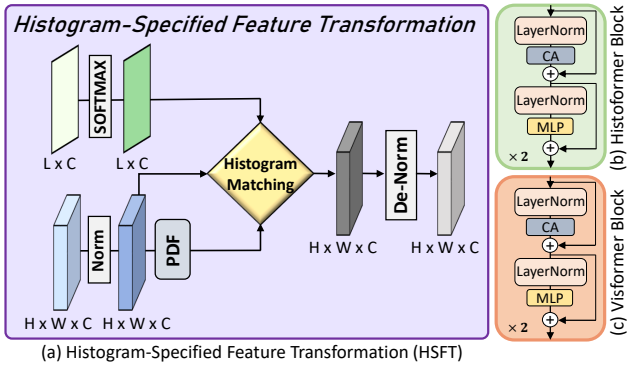

Figure 3: The architecture of (a) Histogram-Specified Feature Transformation Module, (b) Histoformer Block, and (c) Visformer Block.

WB results. First, to train Histoformer, we adopt the L2 loss to calculate the difference between the PDFs of the color channels and those of the ground truth as:

$$\mathcal{L}_{\text{pdf}} = \|\mathbf{H}_{\mathbf{c}}^{\mathbf{R}} - \mathbf{H}_{\mathbf{gt}}^{\mathbf{R}}\|_2 + \|\mathbf{H}_{\mathbf{c}}^{\mathbf{G}} - \mathbf{H}_{\mathbf{gt}}^{\mathbf{G}}\|_2 + \|\mathbf{H}_{\mathbf{c}}^{\mathbf{B}} - \mathbf{H}_{\mathbf{gt}}^{\mathbf{B}}\|_2, \quad (5)$$

where $\mathbf{H_c} = [\mathbf{H}_{\mathbf{c}}^{\mathbf{R}}; \mathbf{H}_{\mathbf{c}}^{\mathbf{G}}; \mathbf{H}_{\mathbf{c}}^{\mathbf{B}}]$ represents the corrected RGB histograms estimated by Histoformer, and likewise $\mathbf{H_{gt}}$ is the ground-truth histograms. Next, we apply the L1 loss to train Visformer as:

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

where $\mathbf{S_c}$ is the estimated WB image by Visformer, and $\mathbf{S_{gt}}$ is the ground truth. The total loss is then defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pdf}} + \mathcal{L}_{\text{rec}}$.

Experiment

Experimental Settings

Training datasets. The Rendered WB Dataset Set1 (Afifi et al. 2019) consists of $62,535$ sRGB images rendered from two public illumination estimation datasets: the NUS dataset (Cheng, Prasad, and Brown 2014) and Gehler dataset (Gehler et al. 2008), using the Adobe Camera Raw feature in Photoshop. The Set1 dataset employs three-fold validation to ensure that the three folds are disjointed with respect to image scenes. Following (Afifi and Brown 2020; Li, Kang, and Ming 2023; Li et al. 2023), we randomly select $12,000$ images from the Set1's fold2 and fold3, which come from various camera models with different WB settings, for training.

Testing sets. To evaluate the performance, we use three commonly used evaluation datasets: Set1-Test ($21,046$ images), Set2 of the Rendered WB Dataset ($2,881$ images), and the sRGB Rendered version of the Cube+ Dataset ($10,242$ images). Set1-Test corresponds to fold1 of the Rendered WB Dataset Set1 (Afifi et al. 2019). The sRGB images in the Rendered WB Dataset Set2 are rendered from raw images of the NUS dataset (Cheng, Prasad, and Brown 2014) by four mobile phones and one DSLR camera, all different from the Set1. It includes $1,874$ DSLR images and $1,007$ mobile phone images. The Rendered Cube+ Dataset consists of $10,242$ images, derived from $1,707$ raw images from the Cube+ Dataset under six illumination conditions (auto correction, as shot, daylight, tungsten, fluorescent, shade) and color-calibrated using a Canon EOS 550D camera across various seasons.

Implementation details. During the training phase, we simultaneously optimize Histoformer and Visformer for 350 epochs. Histoformer is trained using the AdamW opti-

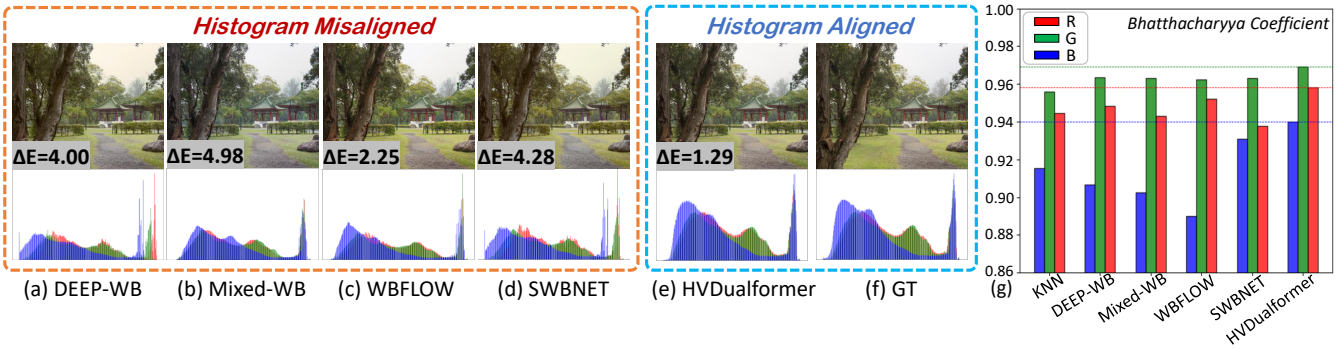

Figure 4: Visualization of global representation in red, green, and blue histograms for existing sRGB-WB methods. 'Histogram Misaligned' indicates that the output images' color histograms are inconsistent with the ground-truth histograms. Our model references corrected color temperature features during the WB process, achieving consistent global representation in color histograms. (g) shows Bhattacharyya Coefficients (Kailath 1967) of color histograms evaluated on the Rendered WB Dataset Set2 (Afifi et al. 2019).

mizer (Loshchilov and Hutter 2017) with a decay rate of the gradient moving average, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, while Visformer is trained using the Adam optimizer (Kingma and Ba 2014) with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ The learning rate is set to $2e - 4$. For performance analysis, we create two versions of the model for further performance analysis, denoted as HVDualformer ($C = 8$) and HVDualformer+ ($C = 12$), where C represents the feature dimension. The model sizes of HVDualformer and HVDualformer+ are 11.05 MB and 24.86 MB, with corresponding parameter counts of 2.92 million and 6.51 million, respectively. During training, we randomly crop four 128×128 patches from the training images as input during training. Additionally, we apply geometric transformations, including rotation and flipping, for data augmentation.

Evaluation metrics. We employ the evaluation metrics as used in (Afifi and Brown 2020; Afifi, Brubaker, and Brown 2022; Li et al. 2023; Li, Kang, and Ming 2023), including Mean Square Error (MSE), Mean Angular Error (MAE), and $\triangle$E 2000 (Sharma, Wu, and Dalal 2005). The reported results of each metric include the mean, lower quartile (Q1), median (Q2), and upper quartile (Q3) of the error.

Experimental Results

We compare HVDualformer with four state-of-the-art (SOTA) WB methods, including five sRGB White Balance methods, KNN (Afifi et al. 2019), DEEP-WB (Afifi and Brown 2020), Mixed-WB (Afifi, Brubaker, and Brown 2022), WBFlow(Li, Kang, and Ming 2023), and SWBNET (Li et al. 2023). All the compared methods were evaluated using their publicly available code and pre-trained models for testing, or their results were cited directly from their respective publications. The exception is SWBNET (Li et al. 2023), which we retrained and tested on the Set1-Test and Set2 datasets, as its original results for these datasets were not provided (marked with $*$ in Table 1). SWBNET's scores on the Cube+ Dataset are taken from the original paper.

Quantitative Results. As presented in Table 1, the quantitative results on the three testing sets, the Rendered WB Dataset Set1-Test, Set2, and the Rendered Cube+ Dataset, demonstrate that the proposed HVDualformer achieves superior performance compared to the compared SOTA methods. Specifically, on Set1-Test and Set2, both versions of our model, HVDualformer and HVDualformer+, generally perform favorably against the competing methods. For the Rendered Cube+ Dataset, HVDualformer+ achieves the best results in mean MSE and mean $\triangle$E 2000 metrics. Although it only shows competitive performance to SWBNET (Li et al. 2023), the best performer in the mean MAE, our model has a much smaller parameter size.

Qualitative Results. Figure 5 and Figure 6 show qualitative comparisons of color correction results on the Rendered WB dataset Set1 and Set2 and Rendered Cube+ Dataset, respectively. As observed, although KNN (Afifi et al. 2019), DEEP-WB (Afifi and Brown 2020), WBFLOW (Li, Kang, and Ming 2023) and Mixed-WB (Afifi, Brubaker, and Brown 2022) can mitigate color casts from the input images to some extent, their performance is not entirely satisfactory. For example, they tend to present slight, unwanted yellowish tones on the walls and in the sky in Figure 5. Similarly, in Figure 6, the wall and building also appear yellowish. In contrast, our method produces results that are natural and color-balanced.

Additionally, Figure 4 compares the color distribution similarity of WB results. Figures 4(a)-(f) show that HVDualformer produces color distributions more closely aligned with the ground truth than other methods. Figure 4(g) further compares the color distribution similarity on the Rendered WB Dataset Set2 (Afifi et al. 2019) using Bhattacharyya Coefficients (Kailath 1967) of red, green, and blue color histograms, which range from 0 to 1 (higher values indicate greater similarity). As shown, HVDualformer achieves the highest similarity to ground truth among all methods.

Ablation Studies

In the ablation studies, we evaluate the effectiveness of different design variations of HVDualformer on the Rendered Cube+ Dataset. Table 2 presents an ablation study with four distinct cases: **Case A** evaluates Histoformer alone, which

Method	MSE				MAE				△E 2000				Size
	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3	MB
Rendered WB Dataset Set1-Test (21,046 images)													
KNN (CVPR 2019)	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 (CVPR 2020)	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 (WACV 2022)	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 (IJCAI 2023)	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* (AAAI 2023)	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
HVDualformer	26.01	6.69	15.29	30.49	2.30°	1.43°	1.99°	2.75°	2.47	1.63	2.19	2.97	11.0
HVDualformer+	**24.09**	**5.67**	**13.61**	**27.99**	**2.14°**	**1.36°**	**1.89°**	**2.59°**	**2.31**	**1.53**	**2.08**	**2.85**	24.8
Rendered WB Dataset Set2 (2,881 images)													
KNN (CVPR 2019)	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 (CVPR 2020)	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 (WACV 2022)	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 (IJCAI 2023)	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* (AAAI 2023)	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
HVDualformer	119.90	29.30	63.36	148.27	3.54°	1.93°	2.82°	4.40°	4.58	3.06	4.08	**5.54**	11.0
HVDualformer+	**110.65**	**26.21**	62.20	147.12	**3.33°**	**1.88°**	**2.70°**	**4.22°**	**4.49**	**3.01**	4.14	5.58	24.8
Rendered Cube+ Dataset with different WB settings (10,242 images)													
KNN (CVPR 2019)	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 (CVPR 2020)	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 (WACV 2022)	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 (IJCAI 2023)	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 (AAAI 2023)	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
HVDualformer	71.73	12.71	31.56	70.16	3.35°	1.68°	2.68°	4.23°	4.25	2.41	3.57	5.25	11.0
HVDualformer+	**68.63**	**11.98**	26.28	**62.65**	3.22°	1.65°	2.50°	**3.99°**	**4.09**	2.31	3.35	**5.02**	24.8

Table 1: WB results on the Rendered WB Dataset (Afifi et al. 2019) and the Rendered version of the Cube+ Dataset (Afifi et al. 2019; Banić, Koščević, and Lončarić 2017)

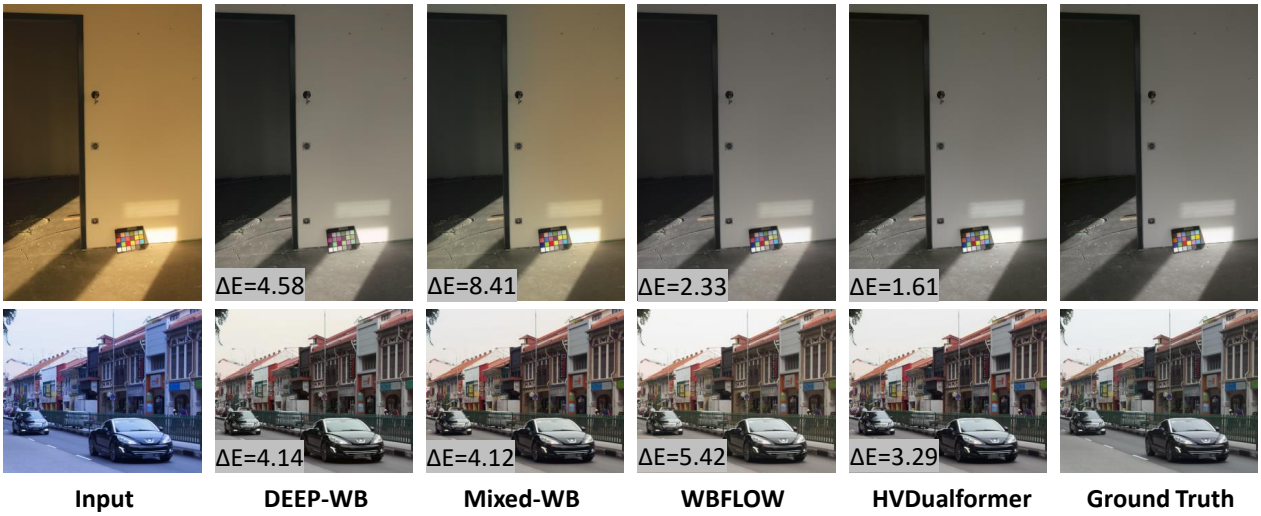

Figure 5: Qualitative comparison of color correction results on the Rendered WB Dataset Set1-Test and Set2 obtained using DEEP-WB (Afifi and Brown 2020), Mixed-WB (Afifi, Brubaker, and Brown 2022), WBFLOW (Li, Kang, and Ming 2023), and the proposed HVDualformer.

performs color correction by aligning the input image with corrected histograms using histogram matching (Coltuc, Bolon, and Chassery 2006) to obtain the color correction results, which achieves only average performance. **Case B** tests Visformer alone, with its parameter size set to $C = 24$ for a fair comparison. Visformer performs better than Histo-

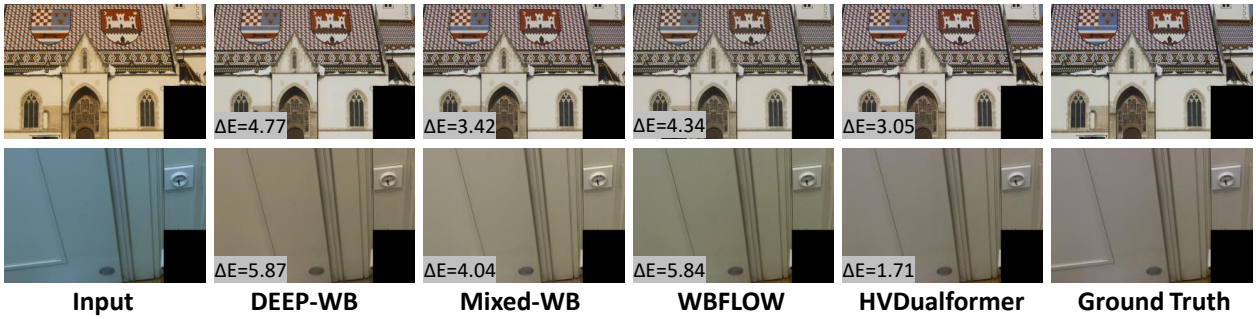

Figure 6: Qualitative comparison of color correction results on the Rendered Cube+ Dataset obtained using DEEP-WB (Afifi and Brown 2020), Mixed-WB (Afifi, Brubaker, and Brown 2022), WBFLOW (Li, Kang, and Ming 2023), and the proposed HVDualformer.

Rendered Cube+ Dataset						
Case	Histoformer	Visformer	HSFT	MSE	MAE	△E
A	✓			168.8	3.84°	5.76
B		✓		115.1	4.08°	4.83
C	✓	✓		138.1	3.59°	4.81
D	✓	✓	✓	**68.6**	**3.22°**	**4.08**

Table 2: Ablation study of different designs in HVDualformer conducted on the Rendered Cube+ Dataset.

Figure 7: Visualization of color correction results from different design cases of our HVDualformer corresponding to Table 2, where Case A is the standalone Histoformer, Case B stands for Visformer, Case C is Histoformer+Visformer w/o HSFT, and Case D is our complete design (HVDualformer).

former. **Case C** connects Histoformer directly to Visformer without the HSFT module. Here, Histoformer adjusts the input image via histogram matching, producing a preliminary result, which is then refined by Visformer. While this approach performs worse than Visformer alone in MSE, it achieves better MAE and △E 2000, making its superiority inconclusive. **Case D** represents our complete design, demonstrating that the proposed HSFT effectively transfers accurate color histograms to remove color casts in the input image. Figure 7 confirms that HVDualformer delivers the most effective color correction among all cases.

Effects of HSFT on Color Correction. The HSFT module modifies and aligns image features using corrected histogram-based color temperature features from Histoformer to ensure balanced color representation. Figure 8 illustrates its effectiveness. As the HVDualformer is designed to learn the residual image features for color correction, we display |**Input-GT**|, the average color channel difference between the input and WB ground truth, along with feature maps before and after HSFT. This comparison shows that HSFT effectively modifies the image features, aligning them more closely with the ground truth residuals, thereby validating its crucial role in our design.

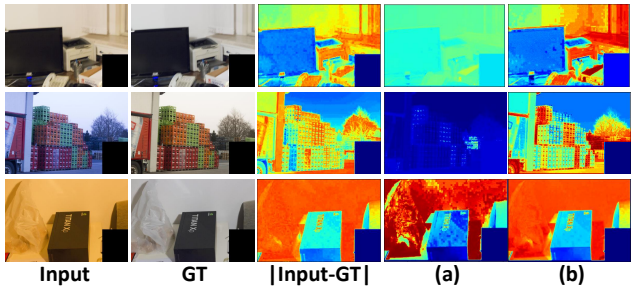

Figure 8: Visualization of HSFT effects on image features. |**Input-GT**| shows the average color channel difference between the input and WB ground truth. Since HVDualformer learns residual image features for color correction, we visualize the feature map (a) before and (b) after HSFT, demonstrating that HSFT effectively aligns features with ground truth residuals.

Conclusion

We proposed HVDualformer, a lightweight histogram-vision dual transformer model for removing color casts from images. It begins with Histoformer, which generates corrected histogram-based color temperature features to adjust input image features through Histogram-Specified Feature Transformation (HSFT). Visformer then processes both the original and HSFT-corrected features to produce the final White Balance (WB) image. Extensive experiments showed that HVDualformer performs favorably against state-of-the-art WB methods both quantitatively and qualitatively.

Acknowledgements

This paper was supported in part by the National Science and Technology Council, Taiwan, under grants NSTC 113-2221-E-004-001-MY3, 113-2622-E-004-001, 113-2221-E-004-006-MY2, 112-2634-F-002-005, 113-2634-F-002-008, and 113-2923-E-A49-003-MY2.

References

Afifi, M.; and Brown, M. S. 2019. What else can fool deep learning? Addressing color constancy errors on deep neural network performance. In *Proc. Int'l Conf. Computer Vision*, 243–252.

Afifi, M.; and Brown, M. S. 2020. Deep white-balance editing. In *Proc. Conf. Computer Vision and Pattern Recognition*.

Afifi, M.; Brubaker, M. A.; and Brown, M. S. 2022. Auto white-balance correction for mixed-illuminant scenes. In *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*.

Afifi, M.; Price, B.; Cohen, S.; and Brown, M. S. 2019. When color constancy goes wrong: Correcting improperly white-balanced images. In *Proc. Conf. Computer Vision and Pattern Recognition*.

Banić, N.; Koščević, K.; and Lončarić, S. 2017. Unsupervised learning for color constancy. *arXiv preprint arXiv:1712.00436*.

Bianco, S.; and Cusano, C. 2019. Quasi-unsupervised color constancy. In *Proc. Conf. Computer Vision and Pattern Recognition*, 12212–12221. Computer Vision Foundation / IEEE.

Brainard, D. H.; and Wandell, B. A. 1986. Analysis of the retinex theory of color vision. *JOSA A*.

Cepeda-Negrete, J.; and Sanchez-Yanez, R. E. 2014. 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*.

Cheng, D.; Prasad, D. K.; and Brown, M. S. 2014. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *JOSA A*.

Coltuc, D.; Bolon, P.; and Chassery, J.-M. 2006. Exact histogram specification. *IEEE Transactions on Image processing*, 15(5): 1143–1152.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gehler, P. V.; Rother, C.; Blake, A.; Minka, T.; and Sharp, T. 2008. Bayesian color constancy revisited. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. IEEE.

Gijsenij, A.; and Gevers, T. 2007. Color constancy using natural image statistics. In *Proc. Conf. Computer Vision and Pattern Recognition*, 1–8. IEEE.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proc. Conf. Computer Vision and Pattern Recognition*.

Hu, Y.; Wang, B.; and Lin, S. 2017. Fc4: Fully convolutional color constancy with confidence-weighted pooling. In *Proc. Conf. Computer Vision and Pattern Recognition*.

Kailath, T. 1967. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kınlı, F.; Yılmaz, D.; Özcan, B.; and Kıraç, F. 2023. Modeling the Lighting in Scenes as Style for Auto White-Balance Correction. In *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*.

Li, C.; Kang, X.; and Ming, A. 2023. WBFlow: Few-shot white balance for sRGB images via reversible neural flows. In *Proc. Int'l Joint Conf. Artificial Intelligence*, 1026–1034.

Li, C.; Kang, X.; Zhang, Z.; and Ming, A. 2023. SWBNet: a stable white balance network for sRGB images. In *Proc. Nat'l Conf. Artificial Intelligence*.

Lo, Y.-C.; Chang, C.-C.; Chiu, H.-C.; Huang, Y.-H.; Chen, C.-P.; Chang, Y.-L.; and Jou, K. 2021. Clcc: Contrastive learning for color constancy. In *Proc. Conf. Computer Vision and Pattern Recognition*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Sharma, G.; Wu, W.; and Dalal, E. N. 2005. 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*.

Shi, W.; Loy, C. C.; and Tang, X. 2016. Deep specialized network for illuminant estimation. In *Proc. Euro. Conf. Computer Vision*.

Van De Weijer, J.; Gevers, T.; and Gijsenij, A. 2007. Edge-based color constancy. *IEEE Trans. on Image Processing*.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proc. Conf. Computer Vision and Pattern Recognition*.