Revisiting Image Fusion for Multi-Illuminant White-Balance Correction Supplementary Material

David Serrano-Lozano[1,2] Aditya Arora[3,4] Luis Herranz[5] Konstantinos G. Derpanis[3,4]
Michael S. Brown[3] Javier Vazquez-Corral[1,2]
[1]Computer Vision Center [2]Universitat Autònoma de Barcelona [3]York University
[4]Vector Institute [5]Universidad Autónoma de Madrid
{dserrano, javier.vazquez}@cvc.uab.cat luis.herranz@uam.es
{adityac8, kosta, mbrown}@yorku.ca

We provide additional material to supplement our main submission, covering:

1. Expanded motivation analysis, further exploring why the ground truth white balance image cannot be obtained by linearly combining WB presets.
2. Additional details about our transformer-based approach.
3. Extended dataset analysis, reporting the ΔE2000, MSE, and MAE metrics of the five white balance (WB) presets, along with additional examples.
4. Additional results.
5. Further ablation studies, examining the type of attention mechanism and the number and selection of WB presets used as input.

1. Expanded Motivation Analysis

Recent sRGB WB methods address multi-illuminant scenes by fusing distinct WB presets, based on the hypothesis that different predefined color temperatures are best suited for regions illuminated by different light sources. MixedWB [3] and StyleWB [6] propose performing the blending as a linear combination. However, the assumption that different illuminants can be merged linearly comes from the RAW linear space but does not hold for non-linear sRGB images.

In the main submission, we empirically demonstrate that, for each pixel, the white-balanced image can lie outside the convex hull formed by the WB presets. Figure 1 presents an additional example from our dataset, showing an image captured under three different illuminants. As before, we select three pixels and plot their values in the sRGB space for the five WB presets (dots) and the ground-truth white-balanced image (crosses). Additionally, we include the corresponding values for MixedWB [3] (diamonds) and our approach (stars). From Figure 1, we can draw two key observations: (i) Any solution derived from linear fusion-based methods will always reside within this convex hull, inherently limiting its ability to reach the optimal correction. (ii) Our transformer-based approach, unconstrained from the linear combination, produces a solution that better approximates the ground-truth image.

2. Additional Method Details

Since linear blending of different WB presets fails short to achieve an optimal white-balanced image, we propose an efficient non-linear fusion approach. We propose using a Transformer Block operating in the feature space [9] and leveraging transposed channel attention [8] to reduce the number of parameters and improve efficiency.

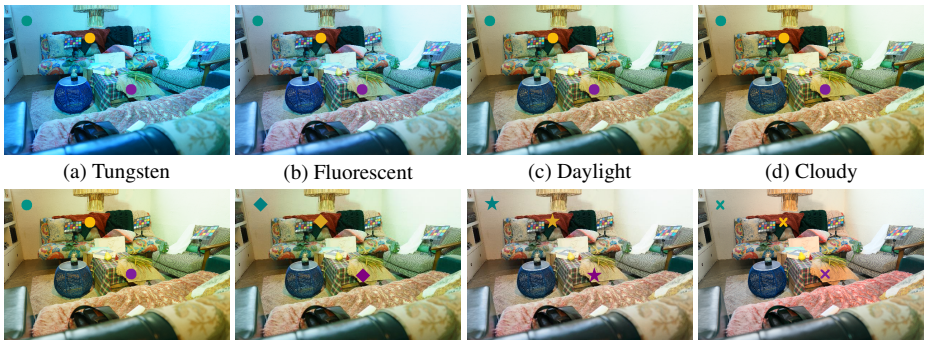
Our approach first renders the RAW sensor image using five predefined WB settings: *tungsten*, *fluorescent*, *daylight*, *cloudy*, and *shade*. These WB-rendered images are concatenated to form a composite image, $I \in \mathbb{R}^{H\times W\times 3P}$, where $H \times W$ represents the spatial dimension of the images and P denotes the number of WB presets. To extract low-level features, we apply a 3×3 convolution, producing the feature maps $F \in \mathbb{R}^{H\times W\times C}$. Following Zhang et al. [9], we generate the *query* (Q), *key* (K), and *value* (V) projections from F by applying a shared *Layer Normalization* followed by independent 1×1 and 3×3 convolutions. This results in three tensors of shape $\mathbb{R}^{H\times W\times C}$, which encode both pixel-wise and channel-wise context. The transposed attention map A is then computed as:

$$A = \text{Softmax}(K^T Q). \tag{1}$$

The transformed features F_a obtained through the Multi-Head Transposed Attention are computed as:

$$F_a = F + W_{1\times 1} AV, \tag{2}$$

where $W_{1\times 1}$ is a point-wise convolution. As in Zamir et al. [8], the channels C are subdivided into multiple heads, allowing parallel learning of separate attention maps.



(a) Tungsten (b) Fluorescent (c) Daylight (d) Cloudy

(e) Shade (f) MixedWB [3] (g) Ours (h) White-balanced image (i) sRGB cube

Figure 1. Example from our dataset. (a)-(e) show the same scene processed with five distinct WB presets, while (f) and (g) present the white-balance result of MixedWB [3] and our transformer-based approach, respectively. Finally, we show the white-balanced image (h). Three sample points, marked with teal, yellow, and purple dots for the WB presets, diamonds for MixedWB [3], stars for our approach, and crosses for the white-balanced image, are selected across all images. (g) visualizes the pixel values in the sRGB space, along with the polytope formed by the WB presets. Note that each axis has a different scale to ease the visualization.

Table 1. Quantitative evaluation of each WB preset on our dataset in terms of ΔE_{00}, MSE, and MAE.. We evaluate the Sony+Nikon configuration.

	ΔE2000			MSE			MAE		
WB Preset	Mean	Median	Trimean	Mean	Median	Trimean	Mean	Median	Trimean
Tungsten (2850K)	21.03	21.47	21.35	283.32	280.41	280.90	22.26	22.02	21.94
Flurorescent (3800K)	15.94	16.15	16.04	266.33	263.42	264.15	13.82	12.96	13.16
Daylight (5500K)	10.82	10.21	10.39	260.63	258.74	259.24	9.46	8.30	8.62
Cloudy (6500K)	11.14	10.47	10.70	261.37	258.91	259.59	10.32	8.89	9.37
Shade(7500K)	12.09	11.68	11.68	262.72	259.79	260.70	11.72	10.23	10.78

Following the attention module, the features are processed by a Feed-Forward Network [5, 7], which enhances contextual information while maintaining per-pixel independence. This second block consists of two parallel paths, each applying a shared *Layer Normalization*, followed by a point-wise and a depth-wise convolutions. The outputs are then fused using a 1×1 convolution. Finally, a 3×3 convolution reduces the channel dimension to three, producing the final sRGB white-balanced image. In our experiment, we set $P = 5$ and $C = 15$.

3. Expanded Dataset Analysis

Our dataset comprises 16,284 sRGB images rendered with five different WB presets: *tungsten* (T), *fluorescent* (F), *daylight* (D), *cloudy* (C), and *shade* (S). In Table 1, we report the ΔE2000, MSE, and MAE for each of the distinct WB presets. The *Daylight* preset consistently achieves the best performance likely due to its neutral color temperature compared to the other WB presets. Please note how the results of our method (Table 1 in the main paper) more than halve the error of this best-preset configuration analysis.

Figure 2 provides additional examples from our dataset. We show three different scenes with three different light setups. For each scene, the first and second rows illustrate two-illuminant setups combining outdoor and indoor light sources. The third row demonstrates a three-illuminant setup, where all light sources are active simultaneously. These examples emphasize how variations in light intensity are accurately captured in the ground truth images, as seen in their brightness levels.

4. Additional Results

Pseudo-weight maps. Unlike linear blending methods that explicitly compute per-pixel weighting maps, our approach estimates the white-balanced image in an end-to-end manner. However, to provide interpretability, we propose generating a set of pseudo-weight maps. These maps are obtained by computing the normalized per-pixel Euclidean distances between the output image and the different WB presets. Figure 3 presents two examples, where the pseudo-weight maps are visualized using the *viridis* color map. In the first scene (also shown in Figure 1), our method produces a white-balanced image that is predominantly closer to the *Daylight* preset. However, certain regions, such as

Table 2. Quantitative evaluation of our method on the Cube+ dataset [4]

	ΔE2000			MSE			MAE		
	Mean	Median	Trimean	Mean	Median	Trimean	Mean	Median	Trimean
Ours	4.19	3.20	3.52	68.33	30.79	37.53	2.98	2.21	2.42

Table 3. Spatial vs. transposed attention ablation study. Results from the combined Sony and Nikon splits of our dataset.

	ΔE2000			MSE			MAE				
	Mean	Med	Trimean	Mean	Med	Trimean	Mean	Med	Trimean	Param	T(ms)
Average Spatial Transformer	4.72	4.52	4.38	92.02	53.40	53.34	4.80	3.45	3.34	105K	248
Concat Spatial Transformer	4.71	4.77	4.59	82.42	50.35	50.91	3.69	3.44	3.37	22.6K	211
Average Transposed Transformer	5.13	5.20	5.02	113.82	77.44	76.09	3.78	3.46	3.43	32.5K	199
Concat Transposed Transformer (ours)	4.55	4.45	4.37	75.60	46.88	49.08	3.61	3.37	3.33	7.9K	179

the specularity on the nearest sofa and the blue table, exhibit stronger similarity to the *Tungsten* and *Fluorescent* presets. This suggests that our approach effectively adapts to local variations in illumination. For the second scene, our method generates a result that aligns more closely with the *Cloudy* preset, demonstrating that is does not overfit to a single WB setting. Additionally, specific image details, such as the backpack and the cushions on the sofa, show greater similarity to the *Tungsten* preset.

Extended Cube+ quantitative results. Due to space limitations in the main submission, Table 2 reports the mean, median and trimean values of ΔE2000, MSE and MAE for our method on the Cube+ dataset [4], trained using a subset of RenderedWB [2].

Qualitative results. Figures 4 and 5 present additional qualitative results from the Nikon and Sony splits of our dataset, respectively. We compare our approach with DeepWB [1] and MixedWB [3]. Notably, our method more effectively removes color casts caused by different illuminants, demonstrating superior performance in achieving accurate white balance.

5. Additional Ablation Studies

Method design. We conduct additional experiments to evaluate the effectiveness of our transposed attention mechanism compared to standard spatial attention. Applying conventional self-attention to high-resolution images is computationally prohibitive, requiring an encoder-decoder structure to reduce feature resolution and attention map size. Thus, we test a model variant that incorporates spatial attention within an encoder-decoder structure. In addition, we explore an alternative method that eliminates the need for the composite image, I, by applying a separate transformer block for each WB preset and averaging their outputs. Table 3 presents the results for the combined Sony and Nikon splits of our dataset, demonstrating that our model (*Concat Transposed Transformer*) consistently achieves superior performance while being the most efficient in terms of parameter count and inference time.

Number and selection of presets. We perform an ablation study to evaluate the impact of the number of presets on performance, demonstrating that five presets yield superior results compared to any combination of just three presets. The study is conducted using the Sony+Nikon configuration of our dataset. The results, shown in Table 4, reveal a significant performance improvement when using five presets, with a 10% increase over the best-performing three-preset configuration.

Table 4. Ablation study for our method when comparing different input presets in terms of ΔE_{00}, MSE, and MAE. Results over the Sony+Nikon split. We can see how using 5 presets is better than any 3 preset configuration.

	ΔE2000			MSE			MAE		
Input presets	Mean	Median	Trimean	Mean	Median	Trimean	Mean	Median	Trimean
D	5.47	3.88	4.02	104.73	75.92	76.04	4.27	3.88	4.02
DST	5.18	4.94	4.96	91.41	62.23	63.03	4.06	3.74	3.76
DFC	5.23	5.22	5.08	93.94	63.50	66.05	4.16	3.97	3.91
DSC	5.23	4.96	4.99	97.57	67.90	69.73	4.34	3.95	3.97
DSF	4.97	4.74	4.75	90.24	63.79	63.54	4.08	3.82	3.77
STC	5.01	4.88	4.84	89.98	59.44	61.68	4.07	3.77	3.75
DSTFC	4.55	4.45	4.37	75.60	46.88	49.08	3.61	3.37	3.33

References

[1] Mahmoud Afifi and Michael S Brown. Deep white-balance editing. In *CVPR*, 2020. 3, 7, 8

[2] Mahmoud Afifi, Brian Price, Scott Cohen, and Michael S Brown. When color constancy goes wrong: Correcting improperly white-balanced images. In *CVPR*, 2019. 3

[3] Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. Auto white-balance correction for mixed-illuminant scenes. In *WACV*, 2022. 1, 2, 3, 7, 8

[4] Nikola Banić, Karlo Koščević, and Sven Lončarić. Unsupervised learning for color constancy. *arXiv preprint arXiv:1712.00436*, 2017. 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[6] Furkan Kınlı, Doğa Yılmaz, Barış Özcan, and Furkan Kıraç. Modeling the lighting in scenes as style for auto white-balance correction. In *WACV*, 2023. 1

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[8] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 1

[9] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 1

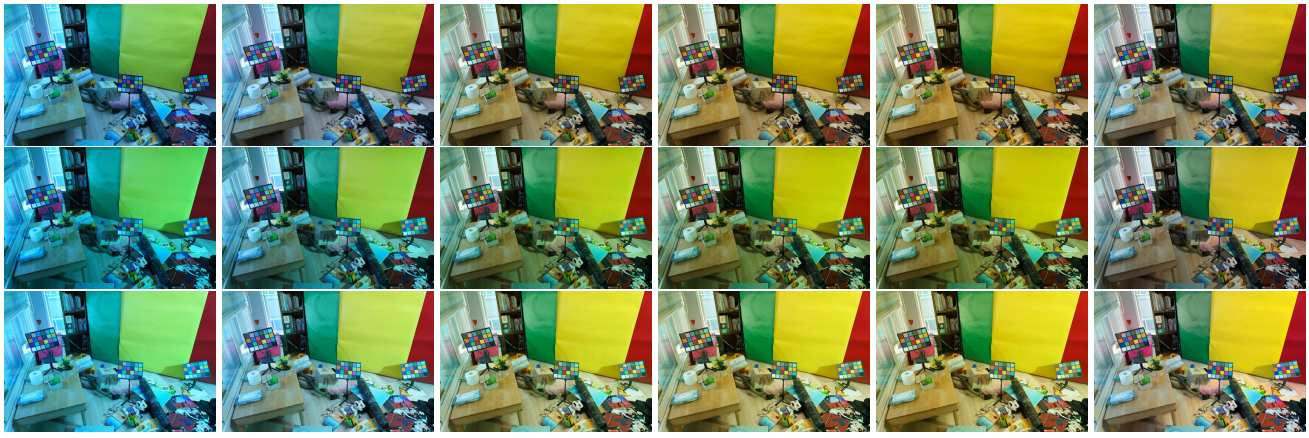


Figure 2. Additional examples from our sRGB dataset for multi-illuminant WB across three scenes. The first and second rows illustrate two illuminant setups, combining outdoor and indoor light sources. The third row shows a three-illuminant setup with all previous light sources active simultaneously. Note how the brightness in the ground truth images reflects the varying light intensity in each scene.

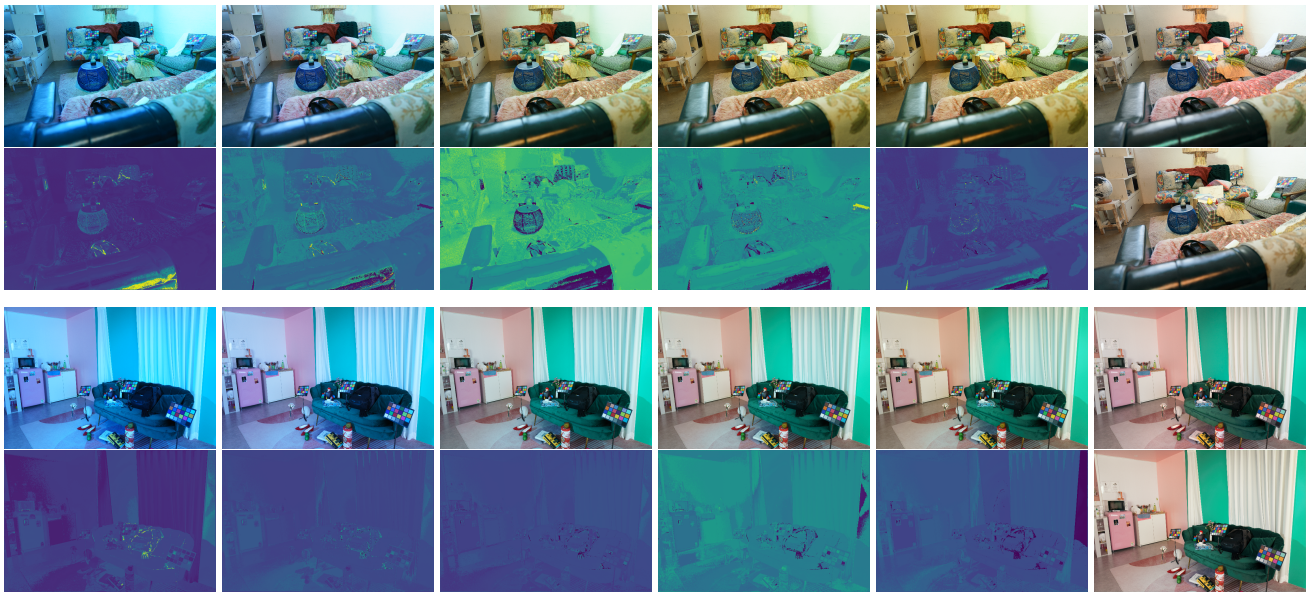
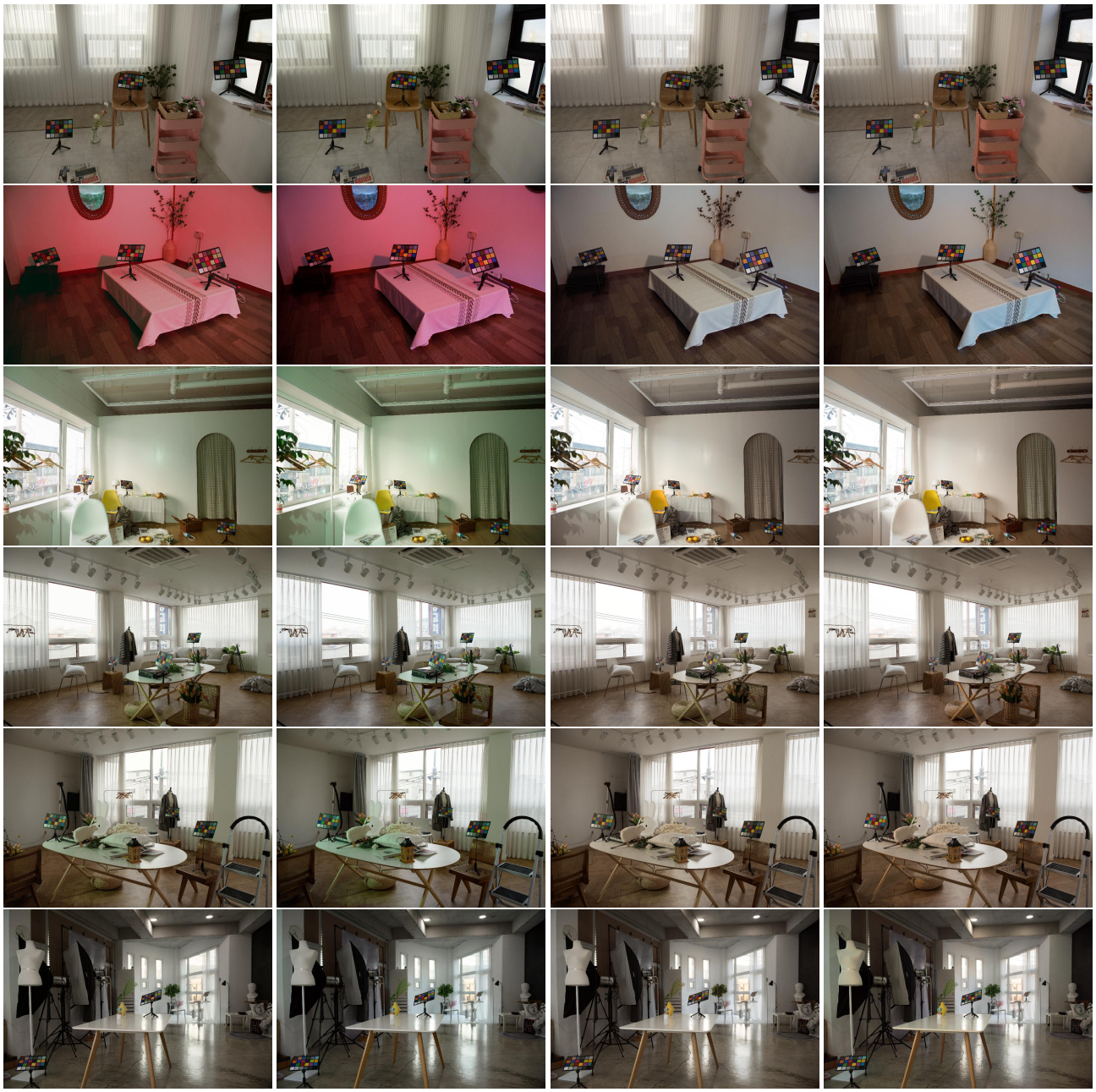


Figure 3. Pseudo-weight maps for two images from our dataset. For each scene, the first row presents the five WB presets alongside the corresponding ground truth. The second row displays the pseudo-weight maps, computed as the normalized per-pixel Euclidean distances between our result and each WB preset. Finally, our white-balanced estimation is shown at the end of the second row.



DeepWB [1] MixedWB [3] Ours Ground truth

Figure 4. Additional results from the Nikon split of our dataset. From left to right, DeepWB [1], MixedWB [3], our transformer-based method, and the ground truth.

Figure 5. Additional results from the Sony split of our dataset. From left to right, DeepWB [1], MixedWB [3], our transformer-based method, and the ground truth.