Supplementary Material
Graph-Based Spectral Attention with Multi-Spectral Images for Illuminant Estimation

Dong-Hoon Kang*
Korea University
Seoul, Korea
rkdehdgns512@korea.ac.kr

Seung-Yeop Baek*
Korea University
Seoul, Korea
sy98baek@korea.ac.kr

Jong-Ok Kim†
Korea University
Seoul, Korea
jokim@korea.ac.kr

1. Complexity Analysis

In this section, we compare the computational complexity of the transformer-based [2] and graph-based attention mechanisms. The transformer-based approach performs attention in both spatial and spectral domains, while ours focuses on spectral feature aggregation in an efficient manner.

1.1. Complexity of Transformer-based Attention

The transformer-based attention mechanism operates in two domains:

- **Spatial attention**: Considers interactions between all patches in an image.
- **Spectral attention**: Models relationships between spectral channels within each patch.

Given an input feature tensor $X \in \mathbb{R}^{c\times h\times w}$, where c is the number of spectral channels and $h \times w$ represents the spatial dimensions, the total computational complexity is:

$$O((hw)^2c + hwc^2) \tag{1}$$

where:

- $O((hw)^2c)$ corresponds to spatial attention, where each patch attends to all others.
- $O(hwc^2)$ accounts for spectral attention, which operates within each patch.

1.2. Complexity of Graph-based Attention

Our graph-based approach performs attention only in the spectral domain while efficiently aggregating spectral features across channels. Instead of computing self-attention in the spatial domain, we utilize a graph structure to model spectral relationships.

Given the adjacency matrix $A \in \mathbb{R}^{c\times c}$, the graph-based operation is formulated as:

$$Z^{(l+1)} = AZ^{(l)}W^{(l)}, \quad l = 0, 1, 2, 3 \tag{2}$$

where $W^{(l)}$ is a trainable weight matrix, and $Z^{(0)} = X$. Since this operation is performed four times, the total computational complexity becomes:

$$O(4hwc^2) \approx O(hwc^2) \tag{3}$$

This formulation eliminates redundant spatial attention computations while retaining spectral feature interactions, leading to improved efficiency and performance.

2. Analysis of Codebook

In this section, we analyze the effect of using codebook in spectral attention mechanisms. Specifically, we compare two cases: (1) when the model uses a codebook with $n_{\text{emb}} = 20$ and (2) when no codebook is used, i.e., $n_{\text{emb}} = 1$. The results demonstrate that utilizing codebook improves spectral feature extraction and enhances model performance.

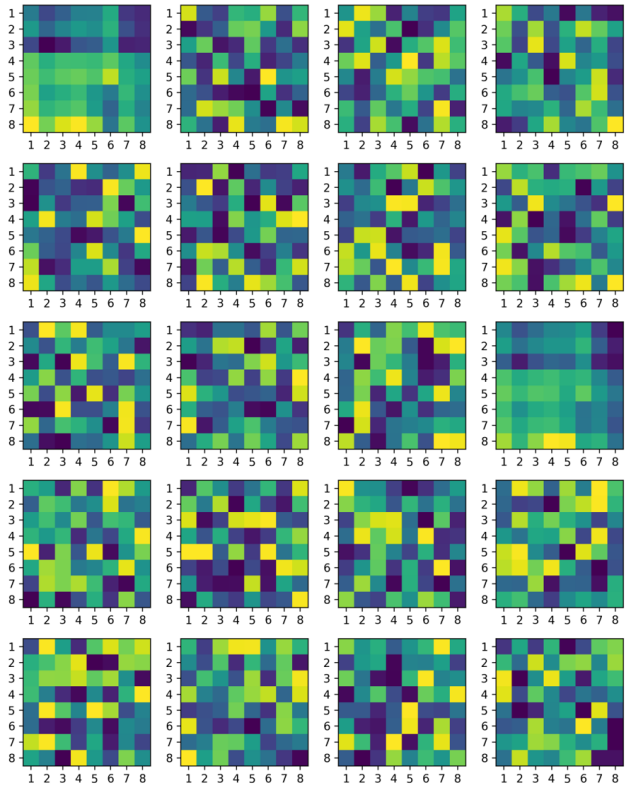
2.1. Limitations of Standard Transformer-based Attention

Standard transformer-based attention mechanisms compute attention scores dynamically based on input features.

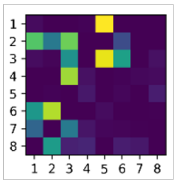
Method	Angular Error			
	Mean	Median	Best-25%	Worst-25%
w/o Codebook	1.51	1.18	0.51	3.03
w/ Codebook	**1.39**	**1.07**	**0.46**	**2.85**

Table 1. Angular error comparison on the NUS-8 dataset with and without the codebook. The use of a codebook improves performance across all metrics, particularly in worst-case scenarios.

*Equal contribution.
†Corresponding author.



(a)



(b)

Figure 1. Visualization of the adjacency matrices with and without codebook. (a) When using codebook, (b) Without codebook.

While this enables adaptability, it also leads to certain limitations:

- **Data Dependency**: Dynamically computed attention scores require large-scale datasets for training to achieve high performance.
- **Computational Cost**: The computation of attention scores for each input leads to high memory and processing overhead.
- **Weak Spectral Correlation Learning**: Since attention scores change based on each input instance, the model fails to effectively capture the underlying spectral correlation across different channels.

On the other hand, graph-based attention mechanisms

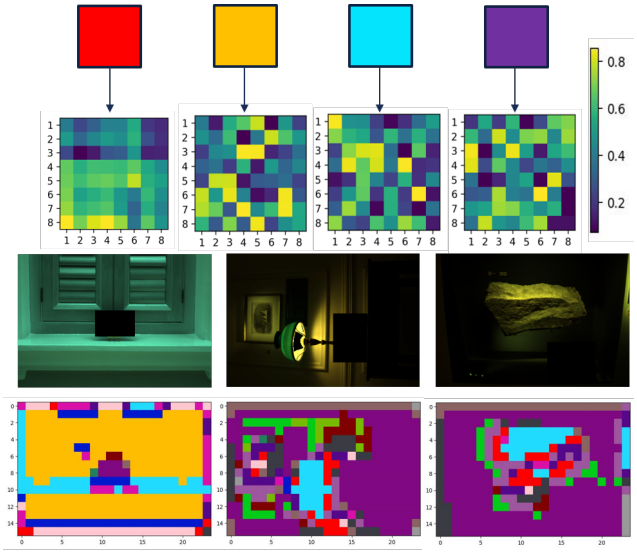


Figure 2. Visualization of patch-wise matrix selection based on brightness and color properties.

leverage a fixed set of learnable matrices to compute attention scores. This approach not only reduces computational cost, but also ensures consistent spectral correlation learning across different inputs.

2.2. Effect of Codebook on Spectral Attention

To address the limitations of dynamic attention scores, we introduce a codebook consisting of predefined matrices that encode spectral correlation information. The effect of the codebook can be analyzed through Fig. 1, which visualizes the matrices in two settings:

- **(a) Codebook with $n_{\text{emb}} = 20$**: A diverse set of 20 matrices, each capturing distinct spectral correlation patterns.
- **(b) No codebook ($n_{\text{emb}} = 1$)**: A single matrix is used universally, failing to capture spectral variations effectively.

Tab. 1 quantifies the performance differences, demonstrating that using a codebook significantly outperforms the case without it. This improvement stems from the model's ability to learn multiple spectral correlation patterns rather than relying on a single matrix.

2.3. Codebook Selection and Spectral Feature Extraction

Fig. 2 illustrates the matrices selected for each patch, revealing an important characteristic of our codebook-based method. Each patch adaptively selects different matrices according to its brightness and color properties. This selection mechanism enables:

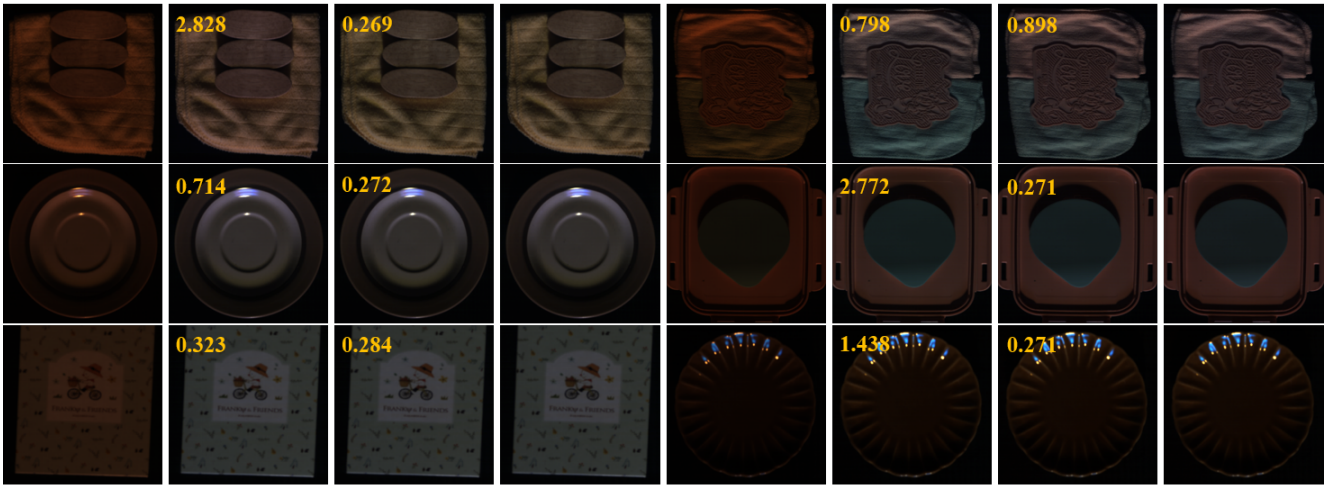


Figure 3. Custom dataset samples. (A), (E): Input RGB. (B), (F): White-balanced with Proposed (RGB). (C), (G): White-balanced with Proposed. (D), (H): White-balanced with GT.

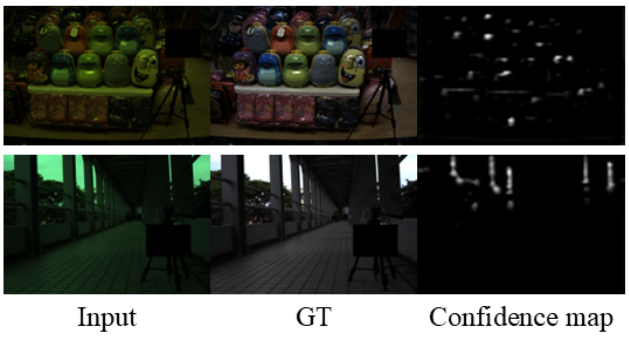


Figure 4. Visualization of high-confidence regions.

- **Brightness and Color Adaptation**: Patches with similar brightness and color tend to select the same matrices, allowing the model to effectively capture spectral correlations related to illumination and material properties.
- **Confidence Map Enhancement**: As shown in Fig. 2, the selection of different matrices based on brightness and color contributes to generating a more reliable confidence map, improving the robustness of illuminant estimation.

The visualization in Fig. 2 confirms that different brightness levels and colors correspond to different matrices, highlighting the model's ability to adaptively extract spectral features while maintaining spectral correlation consistency. This adaptive selection process plays a crucial role in enhancing confidence map generation and overall model performance.

3. Qualitative Analysis

3.1. Ablation Study on Custom Dataset

Fig. 3 illustrates several sample images from our custom dataset, showcasing the effectiveness of our proposed method in illuminant estimation. The numerical values displayed in the top-left corner of each image represent the angular error between the estimated illuminant and the GT illuminant.

We present qualitative results comparing RGB-based illuminant estimation and our proposed MS-assisted approach. The results clearly demonstrate that incorporating spectral information significantly enhances performance across various scenes. Specifically, in challenging cases where the RGB-based method fails to estimate the illuminant accurately due to limited cues or monochromatic surfaces, the MS-assisted approach consistently achieves lower angular errors, leading to improved white balance correction.

3.2. Confidence Map Analysis

To better understand the model's behavior, we analyzed the regions with high confidence in the predicted illumination. We found that the model tends to focus on areas that are originally gray and bright. But, it does not rely on brightness alone; it also reflects spatial context, attending to regions where local pixel distribution contrasts with surroundings, indicating confident regions. This suggests that the model leverages both spectral and spatial cues for illuminant estimation. Examples supporting this observation are shown in Fig. 4.

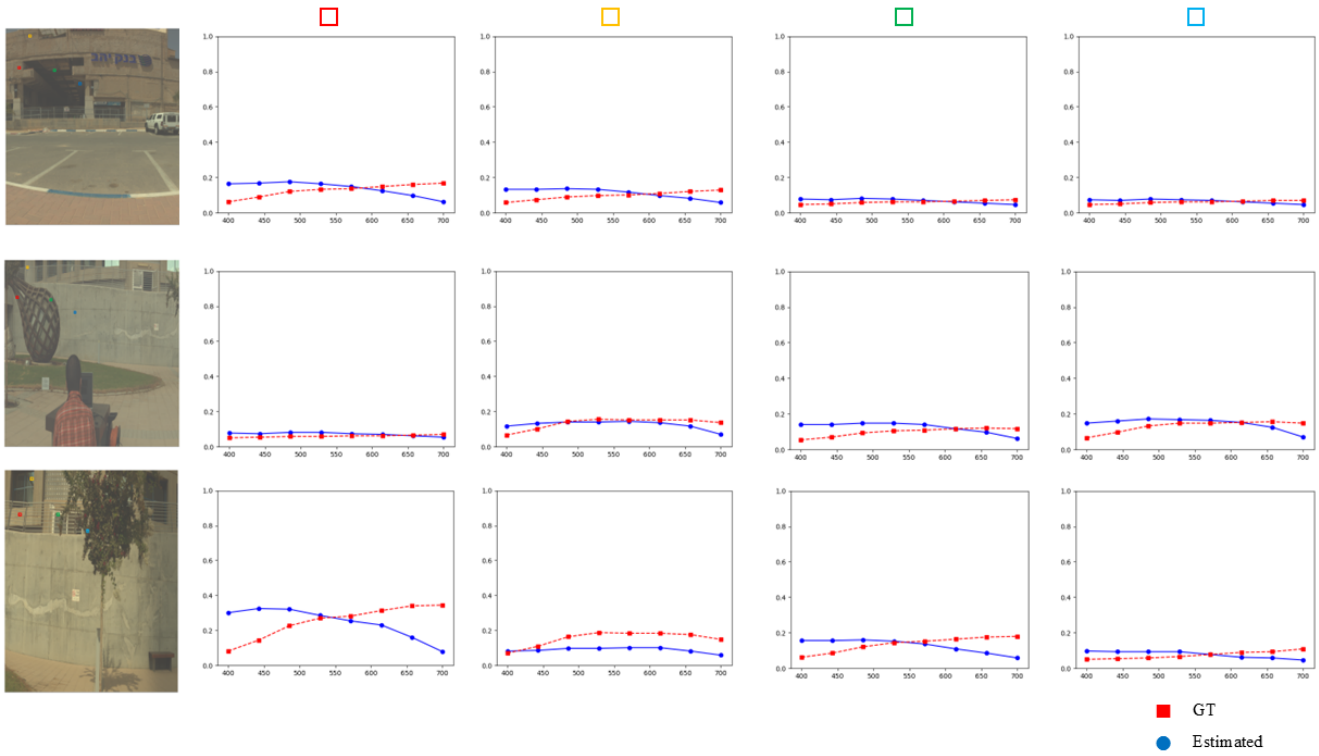


Figure 5. ICVL dataset samples with spectral reflectance comparisons between estimated and GT.

Metric	ICVL Dataset			
	Mean	Median	Best-25%	Worst-25%
PSNR	29.05	28.44	34.37	24.63
SSIM	0.9573	0.9561	0.9715	0.9402

Table 2. PSNR and SSIM evaluation on the ICVL dataset. The table presents mean, median, best 25%, and worst 25% values.

4. Evaluation of RGB-to-MS Conversion

Fig. 5 presents several ICVL dataset samples that were not included in the training set for the RGB-to-MS conversion model. Each corresponding graph illustrates the spectral reflectance of selected pixels, allowing for a direct comparison between the estimated reflectance and the GT reflectance.

As observed in the graphs, the estimated spectral reflectance closely follows the GT values across different wavelengths. Furthermore, Tab. 2 quantitatively evaluates the accuracy of the estimated MS images by comparing their PSNR and SSIM with the GT MS images. The high PSNR and SSIM values indicate that even though the MS images are estimated from RGB inputs, they retain spectral characteristics that are highly similar to the GT MS data.

These results confirm the effectiveness of the RGB-to-MS conversion model in preserving spectral information, reinforcing its applicability for downstream tasks that require spectral fidelity.

5. Discussion with Relevant Method [1]

[1] is highly relevant to our study, but it states that the official codes and dataset are not publicly available. Therefore, we analyzed it in a discussion-based manner rather than through direct experimental comparison.

[1] models the actual physical sensor responses using Gaussian functions, and the modeled sensors are then used to estimate the spectral sensor responses from RGB images. Physically modeling the sensor responses is part of their contribution. Our model adapts the NIR-to-RGB conversion model from [3], with RGB applied instead of NIR.

Also, [1] deliberately does not use spatial information, and this is guessed as part of their novelty. On the other hand, our method estimates the illuminant primarily from spatial features extracted from RGB images, while using spectral features to predict the confidence of each local illuminant estimation.

While [1] employs the estimated spectra with 14 and 65 channels, we use 8-channel estimated MS images. More channels can potentially provide richer information and improve performance, but they also increase computational cost. We chose 8 channels as a practical trade-off between

performance and efficiency.

References

[1] Samu Koskinen, Erman Acar, and Joni-Kristian Kämäräinen. Single pixel spectral color constancy. *International Journal of Computer Vision*, 132(2):287–299, 2024. 4

[2] Miaoyu Li, Ying Fu, and Yulun Zhang. Spatial-spectral transformer for hyperspectral image denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1368–1376, 2023. 1

[3] Tae-Sung Park, Tae-Hyeon Kim, and Jong-Ok Kim. Feature distillation network for multi-band nir colorization. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1874–1878. IEEE, 2022. 4