Supplementary Material of:
Auto White-Balance Correction for Mixed-Illuminant Scenes

Mahmoud Afifi[1] Marcus A. Brubaker[1,2] Michael S. Brown[1,2]

[1]York University [2]Vector Institute

{mafifi,mab,mbrown}@eecs.yorku.ca

This supplementary material includes details of our network architecture (Sec. 1) and our synthetic testing set (Sec. 2), additional ablation studies (Sec. 3), and additional qualitative results (Sec. 4).

1. Network Architecture

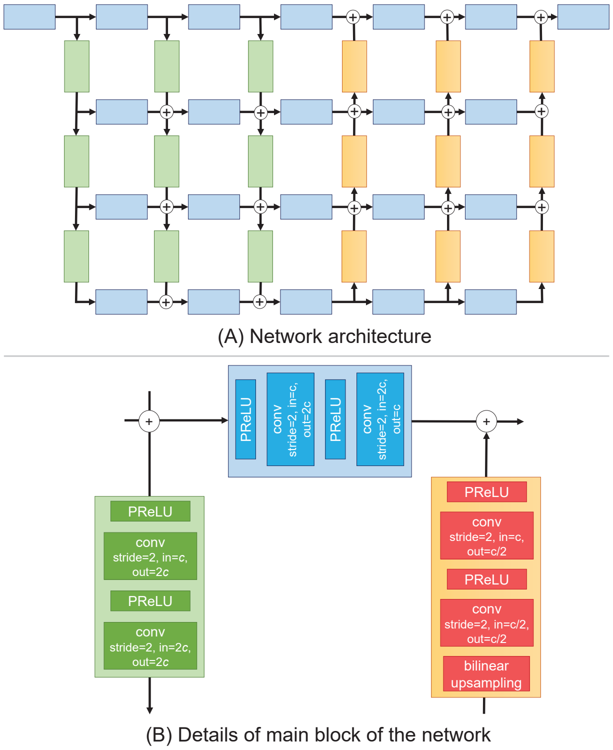
We adopted the GridNet architecture [11, 15]. Our network consists of six columns and four rows. As shown in Figure 1, our network includes three main units, which are: the residual unit (shown in blue in Figure 1), the downsampling unit (shown in green in Figure 1), and the upsampling unit (shown in yellow in Figure 1).

The number of input/output channels, stride, and padding size of each conv layer are shown in Figure 1-(B). The first residual unit of our network accepts concatenated input images with 3×k channels, where k refers to the number of images rendered with k WB settings. For example, when using WB={$\texttt{t}, \texttt{d}, \texttt{s}$}, the value of k is three. Regardless of the value of k, we set the number of output channels of the first residual to unit to eight.

Each residual block (except for the first one), produces features with the same dimensions of the input feature. For the first three columns, the dimensions of each feature received from the upper row are reduced by two, while the number of output channels is duplicated, as shown in the downsampling unit in Figure 1-(B). In contrast, the upsampling unit (shown in Figure 1-[B]) increases the dimensions of the received features by two in the last three columns. Lastly, the last residual unit produces output weights with k channels.

2. Our Synthetic Test Set

As mentioned in the main paper, we have generated a set of 150 images with mixed illuminations. The ground-truth of each image is provided by rendering the same scene with a fixed color temperature used for all light sources in the scene and the camera AWB.


(A) Network architecture

PReLU
conv stride=2, in=c, out=2c
PReLU
conv stride=2, in=2c, out=c
PReLU
conv stride=2, in=c, out=2c
PReLU
conv stride=2, in=2c, out=2c
PReLU
conv stride=2, in=c, out=c/2
PReLU
conv stride=2, in=c/2, out=c/2
bilinear upsampling

(B) Details of main block of the network

Figure 1: Network architecture. We adopted the GridNet architecture [11, 15], as shown in (A). (B) shows the details of the residual, downsampling, and upsampling units. The symbol c refer to the number of channels in each conv layer.

Existing paired multi-illuminant datasets provide ground-truth images as either color maps or albedo layers (i.e., reflectance) [9, 8, 14, 13]. For instance, the two-illuminant dataset [8] includes 78 images (58 laboratory images taken under close-to-ideal conditions and 20 real-world images). Each test image has a corresponding ground-truth illuminant color map, as shown in Figure 2-(B). Unfortunately, the two-illuminant dataset [8] cannot

Table 1: Impact of ensembling and edge-aware smoothing (EAS) [6] at inference time. In this set of experiments, we used WB=$\{\texttt{t}, \texttt{d}, \texttt{s}\}$ with training patch-size $p = 64$. We reported the mean, first, second (median), and third quantile (Q1, Q2, and Q3) of mean square error (MSE), mean angular error (MAE), and $\triangle$**E** 2000 [18]. The top results are indicated with yellow and bold.

Method	MSE				MAE				△E 2000			
	Mean	**Q1**	**Q2**	**Q3**	**Mean**	**Q1**	**Q2**	**Q3**	**Mean**	**Q1**	**Q2**	**Q3**
w/o ensembling, w/o EAS	849.33	694.68	846.80	1051.24	5.58°	4.55°	5.05°	6.46°	10.73	9.51	10.76	11.95
w/ ensembling, w/o EAS	831.41	662.25	855.24	1005.04	**5.42°**	4.38°	4.98°	**6.08°**	10.64	9.46	**10.66**	11.91
w/ ensembling, w/ EAS	**819.47**	**655.88**	**845.79**	**1000.82**	5.43°	**4.27°**	**4.89°**	6.23°	**10.61**	**9.42**	10.72	**11.81**

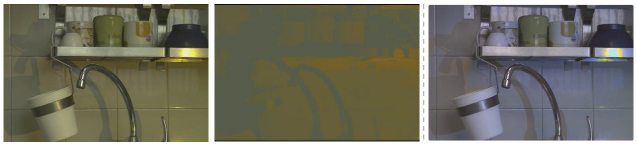

(A) Input image (B) Illuminant map (D) Ground-truth image

Figure 2: Example from the two-illuminant dataset [8]. (A) Input sRGB image. (B) Provided ground-truth illuminant map. (C) Generated ground-truth image using the illuminant map in (B).

be used by our method as it does not provide the same scene rendered with different WB settings. The provided raw images are in the PNG format and there are no DNG metadata provided to render such images to sRGB with different WB settings. In addition, creating the final ground-truth image in sRGB space, given the ground-truth illuminant map, does not always give satisfying results; see the specular region shown in Figure 2-(C).

Recently, the MIST dataset [14, 13] was generated by rendering 3D scenes using Blender Cycles [1]. The MIST dataset was proposed to handle the limitations of existing multi-illuminant datasets (e.g., [8, 12, 7]) by providing accurate ground-truth image intrinsic properties (i.e., albedo, diffuse, and specular layers) after an accurate measurement of the illumination at every point in each 3D scene. Despite its useful impact on testing image intrinsic decomposition methods, there is no ground-truth image provided for our task—namely, pixel-wise image white balancing.

Due to the aforementioned limitations of existing multi-illuminant test sets, we generated our test set with accurate ground-truth images for evaluating WB methods targeting mixed-illuminant scenes; see Figure 3

3. Additional Ablation Studies

In the main paper, we showed the results of ablation studies on the impact of different settings, including the size of training patches (p), the WB settings used to render input small images, and the smoothing loss term ($\mathcal{L}_s$).

In this section, we show additional ablation study conducted to show the impact of the ensembling step and the post-processing edge-aware smoothing (EAS) step [6]. Table 1 shows the results of our method with and without the ensemble approach and the EAS post-processing step. The size of the input images is 384×384 pixels when ensembling is applied; otherwise, we used images of 256×256 pixels. Empirically, we found that image size of 256×256 pixels or 128×128 pixels give always the best results when the ensemble approach is not used.

Figures 4 and 5 show qualitative comparisons of our results with and without the ensembling and the EAS steps. As shown, when the ensemble testing is used, our predicted weights have more local coherence, which is further improved when using the EAS step.

4. Additional Results

In the main paper, we reported quantitative results on our test set of our method and other methods for WB correction, which are: gray pixel [16], grayness index [17], KNN WB [5], Interactive WB [4], and Deep WB [3]. Here, we show qualitative comparisons between our method and the aforementioned methods in Figure 6.

Finally, we show additional results on the MIT-Adobe 5K dataset [10] in Figure . Note that none of cameras/images in these sets were used in training either our method or the other methods.

References

[1] Blender. `https://www.blender.org`. Accessed: 2021-07-20.

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

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

[4] Mahmoud Afifi and Michael S Brown. Interactive white balancing for camera-rendered images. In *Color and Imaging Conference*, 2020.

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

[6] Jonathan T Barron and Ben Poole. The fast bilateral solver. In *ECCV*, 2016.

[7] Shida Beigpour, Mai Lan Ha, Sven Kunz, Andreas Kolb, and Volker Blanz. Multi-view multi-illuminant intrinsic dataset. In *BMVC*, 2016.

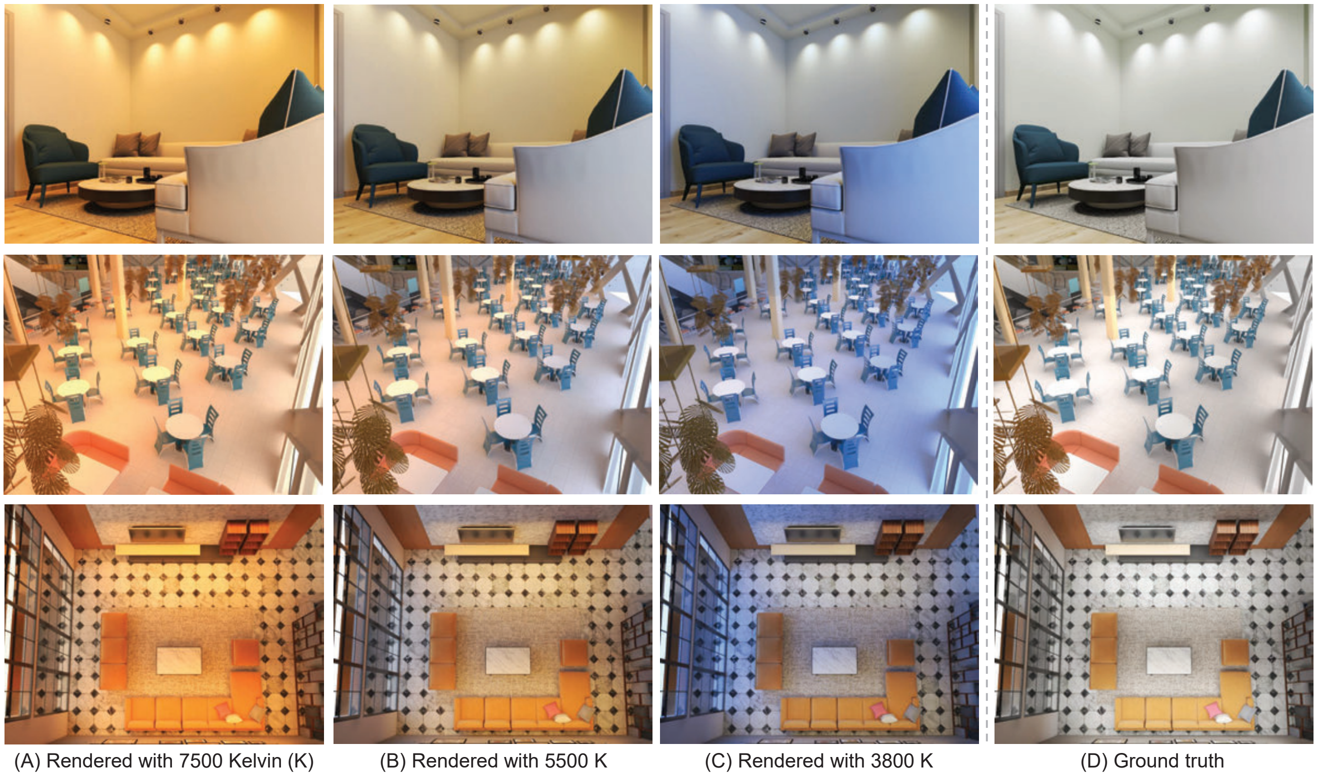


Figure 3: Examples from our synthetic test set. (A-C) Rendered images with different color temperatures that are associated to the following WB settings: shade, daylight, and fluorescent, respectively [2]. (D) Ground-truth image.

[8] Shida Beigpour, Christian Riess, Joost Van De Weijer, and Elli Angelopoulou. Multi-illuminant estimation with conditional random fields. *IEEE Transactions on Image Processing*, 23(1):83–96, 2013.

[9] Michael Bleier, Christian Riess, Shida Beigpour, Eva Eibenberger, Elli Angelopoulou, Tobias Tröger, and André Kaup. Color constancy and non-uniform illumination: Can existing algorithms work? In *ICCV Workshops*, 2011.

[10] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *CVPR*, 2011.

[11] Damien Fourure, Rémi Emonet, Elisa Fromont, Damien Muselet, Alain Tremeau, and Christian Wolf. Residual conv-deconv grid network for semantic segmentation. In *BMVC*, 2017.

[12] Arjan Gijsenij, Rui Lu, and Theo Gevers. Color constancy for multiple light sources. *IEEE Transactions on Image Processing*, 21(2):697–707, 2011.

[13] Xiangpeng Hao and Brian Funt. A multi-illuminant synthetic image test set. *Color Research & Application*, 45(6):1055–1066, 2020.

[14] Xiangpeng Hao, Brian Funt, and Hanxiao Jiang. Evaluating colour constancy on the new mist dataset of multi-illuminant scenes. In *Color and Imaging Conference*, volume 2019, 2019.

[15] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *CVPR*, 2018.

[16] Yanlin Qian, Ke Chen, Jarno Nikkanen, Joni-Kristian Kämäräinen, and Jiri Matas. Revisiting gray pixel for statistical illumination estimation. In *VISAPP*, 2019.

[17] Yanlin Qian, Joni-Kristian Kämäräinen, Jarno Nikkanen, and Jiri Matas. On finding gray pixels. In *CVPR*, 2019.

[18] Gaurav Sharma, Wencheng Wu, and Edul N Dalal. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005.

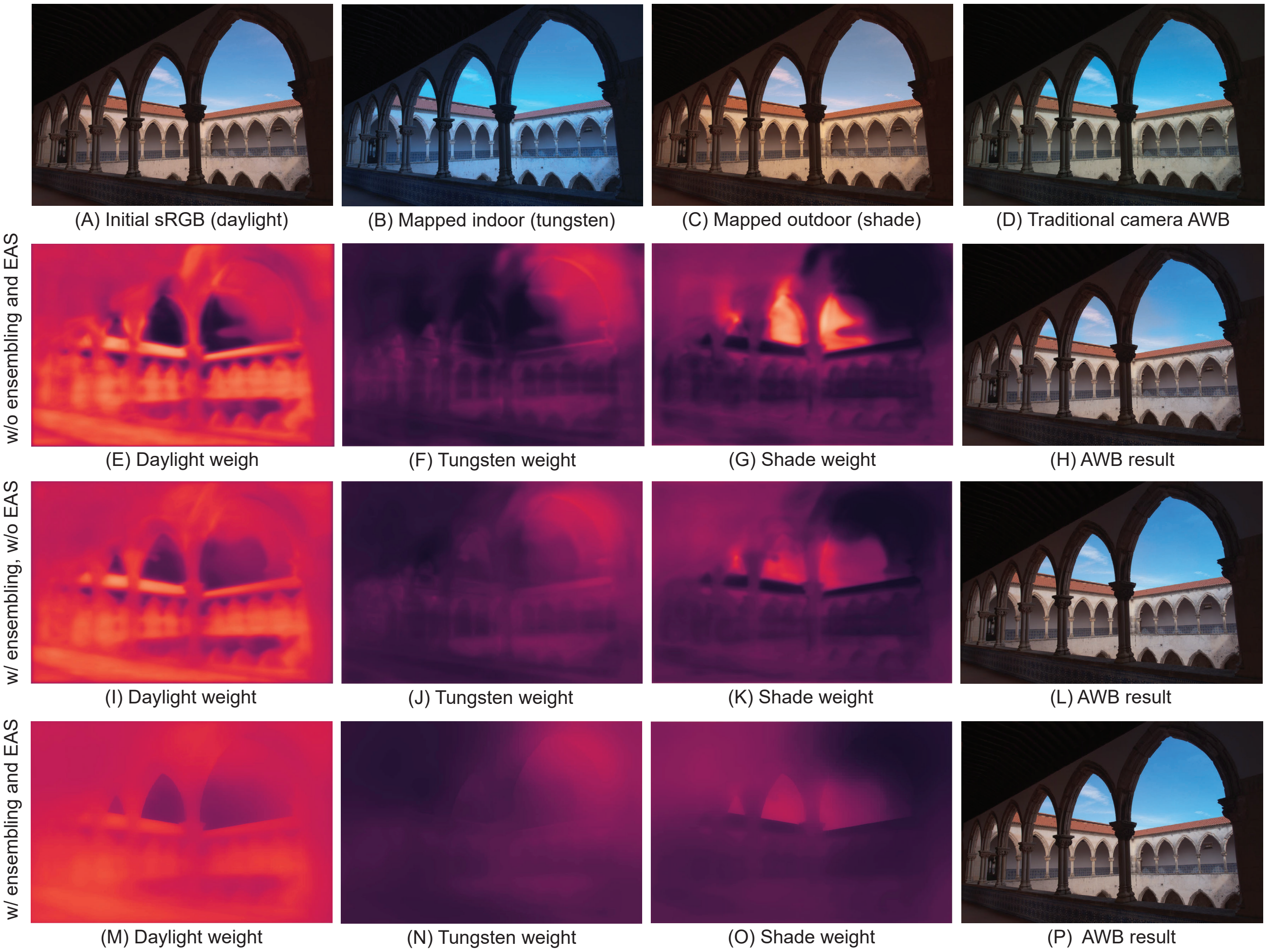


Figure 4: Qualitative examples showing the impact of ensembling and edge-aware smoothing (EAS) at inference time. The first row shows: (A) the initial rendered image with daylight white-balance setting, (B-C) indoor and outdoor high-resolution images after mapping, and (D) the result of traditional camera AWB correction. The second row shows: (E-G) the predicted weights without ensembling nor the EAS post-processing, along with the final AWB result after blending in (H). The third row shows the results when using the ensemble processing. The fourth row shows the results when using ensembling and EAS post-processing. Input images are from the the MIT-Adobe 5K dataset [10].

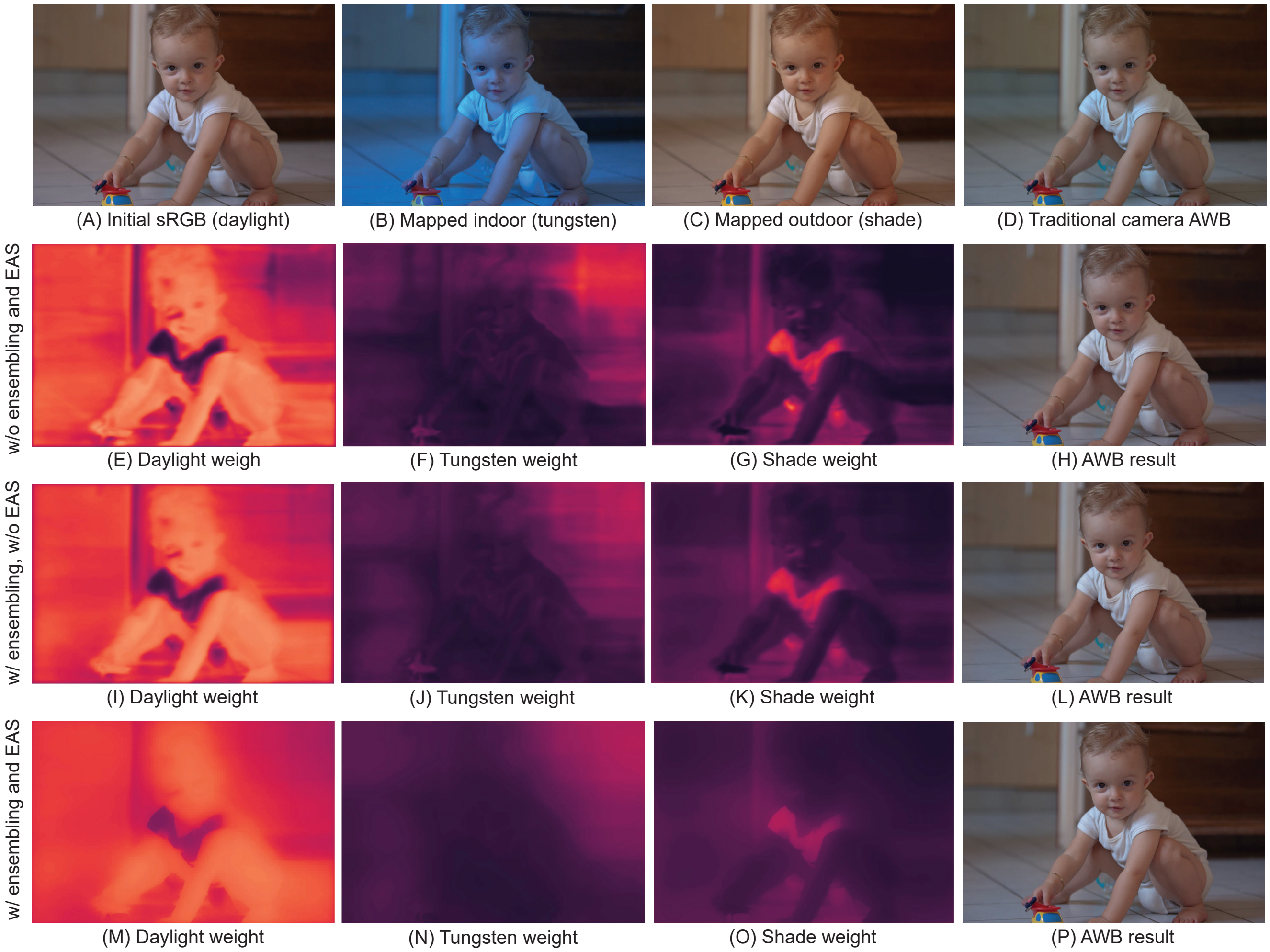
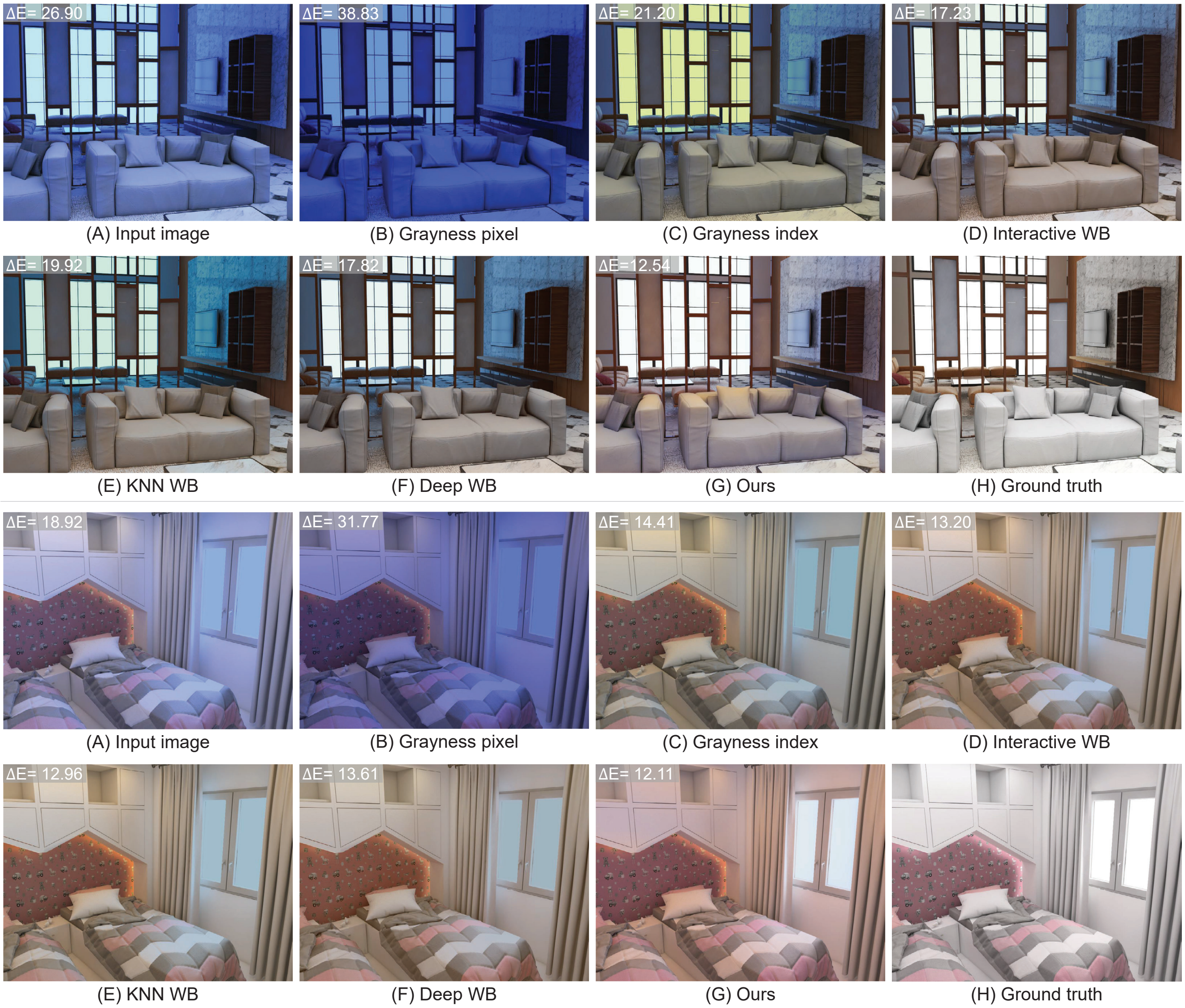


Figure 5: Additional qualitative examples showing the impact of ensembling and edge-aware smoothing (EAS) at inference time. The first row shows: (A) the initial rendered image with daylight white-balance setting, (B-C) indoor and outdoor high-resolution images after mapping, and (D) the result of traditional camera AWB correction. The second row shows: (E-G) the predicted weights without ensembling nor the EAS post-processing, along with the final AWB result after blending in (H). The third row shows the results when using the ensemble processing. The fourth row shows the results when using ensembling and EAS post-processing. Input images are from the the MIT-Adobe 5K dataset [10].



(A) Input image (B) Grayness pixel (C) Grayness index (D) Interactive WB

(E) KNN WB (F) Deep WB (G) Ours (H) Ground truth

(A) Input image (B) Grayness pixel (C) Grayness index (D) Interactive WB

(E) KNN WB (F) Deep WB (G) Ours (H) Ground truth

Figure 6: Qualitative comparisons with other AWB methods on our mixed-illuminant evaluation set. Shown are the results of the following methods: gray pixel [16], grayness index [17], interactive WB [4], KNN WB [5], deep WB [3], and our method.

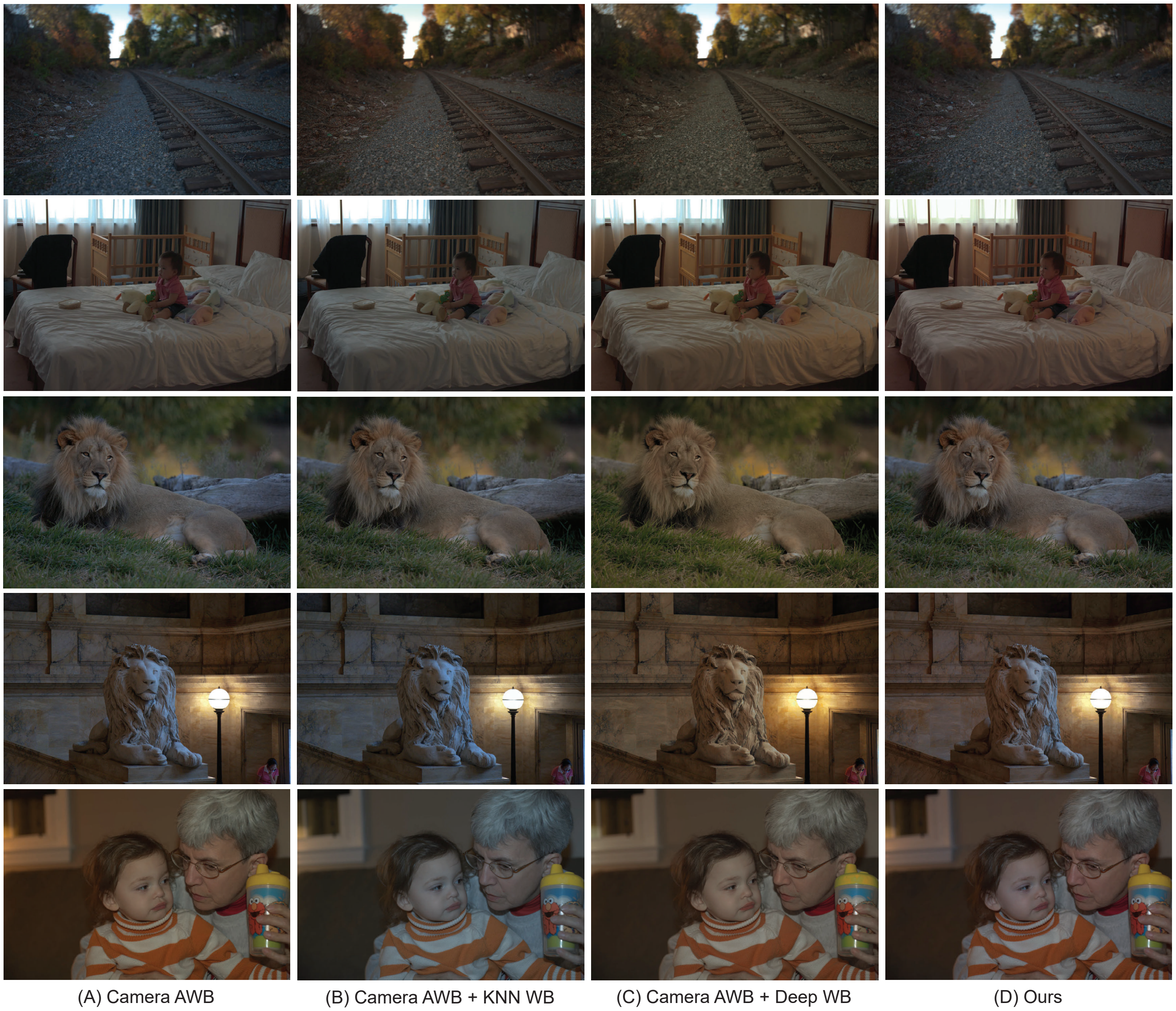


Figure 7: Additional qualitative comparisons with other AWB methods on the MIT-Adobe 5K dataset [10]. Shown are the results of the following methods: gray pixel [16], grayness index [17], interactive WB [4], KNN WB [5], deep WB [3], and our method.