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

Supplementary Material

1. Overview

This supplementary material provides additional resources to complement the main manuscript, including the following:

- A visualization of the network architecture used to extend the proposed ABC-Former for handling images captured under multiple light sources.
- Quantitative and qualitative evaluations conducted on the Mixed-illuminant dataset [3].
- Additional qualitative results from various datasets, including the MIT-Adobe 5K dataset [6], the Rendered WB dataset (Set1-Test and Set2) [2], and the Rendered Cube+ dataset [2, 4].

2. Extension to the Multi-illuminant Task

Handling color casts caused by multiple light sources within a single scene presents a significant challenge for single-illuminant white balance (WB) methods. To address it, recent studies [3, 8] have proposed generating blending weighting maps instead of estimating a single illuminant for the entire scene. These maps correspond to the various WB settings required to adapt to scenes illuminated by multiple light sources.

Our proposed model, initially designed for single-illuminant WB tasks, can be easily extended to the multi-illuminant WB task by leveraging the original ABC-Former as its backbone. This extended version, termed **ABC-FormerM**, differs from the standard ABC-Former only regarding its inputs and outputs. Specifically, ABC-FormerM takes multi-illuminant images as input and generates weighting maps corresponding to each input image. Figure 1 depicts the architecture of ABC-FormerM, using an example with three WB settings: "tungsten," "daylight," and "shade." The input images are first concatenated and passed through a 1×1 convolutional layer to reduce their channels to three for processing by the sRGBformer. Similarly, the sRGB and CIELab histograms of the input images are concatenated and convolved with a 1×1 kernel to prepare them for PDFformers, respectively. Unlike ABC-Former, which directly outputs a WB-corrected image, ABC-FormerM generates weighting maps for each multi-illuminant image. These maps are subsequently refined using edge-aware smoothing via a fast bilateral solver [5]. The final WB-corrected image is obtained by applying multi-scale weighted averaging to the multi-illuminant images, following the generation method described in [3, 8].

2.1. Loss Function

To optimize ABC-FormerM, we employ the original loss functions from **Section 3.3: Loss Function** of the main manuscript to supervise the auxiliary models ($\mathcal{L}_{\text{pdf}}^{sRGB}$ and $\mathcal{L}_{\text{pdf}}^{Lab}$). Additionally, we incorporate the reconstruction loss ($\mathcal{L}_{\text{rec}}$) and smoothing loss ($\mathcal{L}_{\text{smooth}}$), as following the approaches used in Mixed-WB [3] and Style WB [8], defined respectively as:

$$\mathcal{L}_{\text{rec}} = \left\| \mathrm{P}_{\text{gt}} - \sum_i \hat{W}_i \mathrm{P}_i \right\|_F^2; \tag{1}$$

$$\mathcal{L}_{\text{smooth}} = \sum_i \left\| \hat{W}_i * \nabla_x \right\|_F^2 + \left\| \hat{W}_i * \nabla_y \right\|_F^2, \tag{2}$$

where $\mathbf{P_{gt}}$ refers to the ground truth patch, and $\mathbf{P_i}$ represents the input patch rendered with the i_{th} WB setting. The corresponding weighting map for $\mathbf{P_i}$ is denoted as $\hat{W}_i$, where $i \in \{t, d, s\}$ or $i \in \{t, f, d, c, s\}$. These initials represent the WB settings: ***t***ungsten, ***f***luorescent, ***d***aylight, ***c***loudy, and ***s***hade, respectively. The smoothing loss $\mathcal{L}_{\text{smooth}}$ is computed by the weighting map convolved with 3×3 horizontal and vertical Sobel filters, ∇_x and ∇_y. The total loss function is defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pdf}}^{sRGB} + \mathcal{L}_{\text{pdf}}^{Lab} + \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{smooth}}$, with λ set to 100.

2.2. Evaluation on the Multi-illuminant Conditions

We conducted experiments to evaluate the performance of ABC-FormerM on the mixed-illumination task. The experiments were performed using two different patch sizes (64×64 and 128×128) and two sets of predefined WB settings: (i) $\{t, d, s\}$, and (ii) $\{t, f, d, c, s\}$. These settings correspond to the following color temperatures: tungsten (2850 K), fluorescent (3800 K), daylight (5500 K), cloudy (6500 K), and shade (7500 K). All other experimental configurations, including the learning rate and optimizer, matched those used in the original ABC-Former.

We evaluated the model on the Mixed-illuminant dataset [3], which contains 150 synthetic images with multiple light sources created from 3D scenes modeled in Autodesk 3Ds Max. ABC-FormerM was compared against state-of-the-art WB methods, including single-illuminant WB methods [1, 2, 9], as well as multi-illuminant WB methods [3, 8]. Table 1 reports the WB performance using three objective metrics: Mean Square Error (MSE),

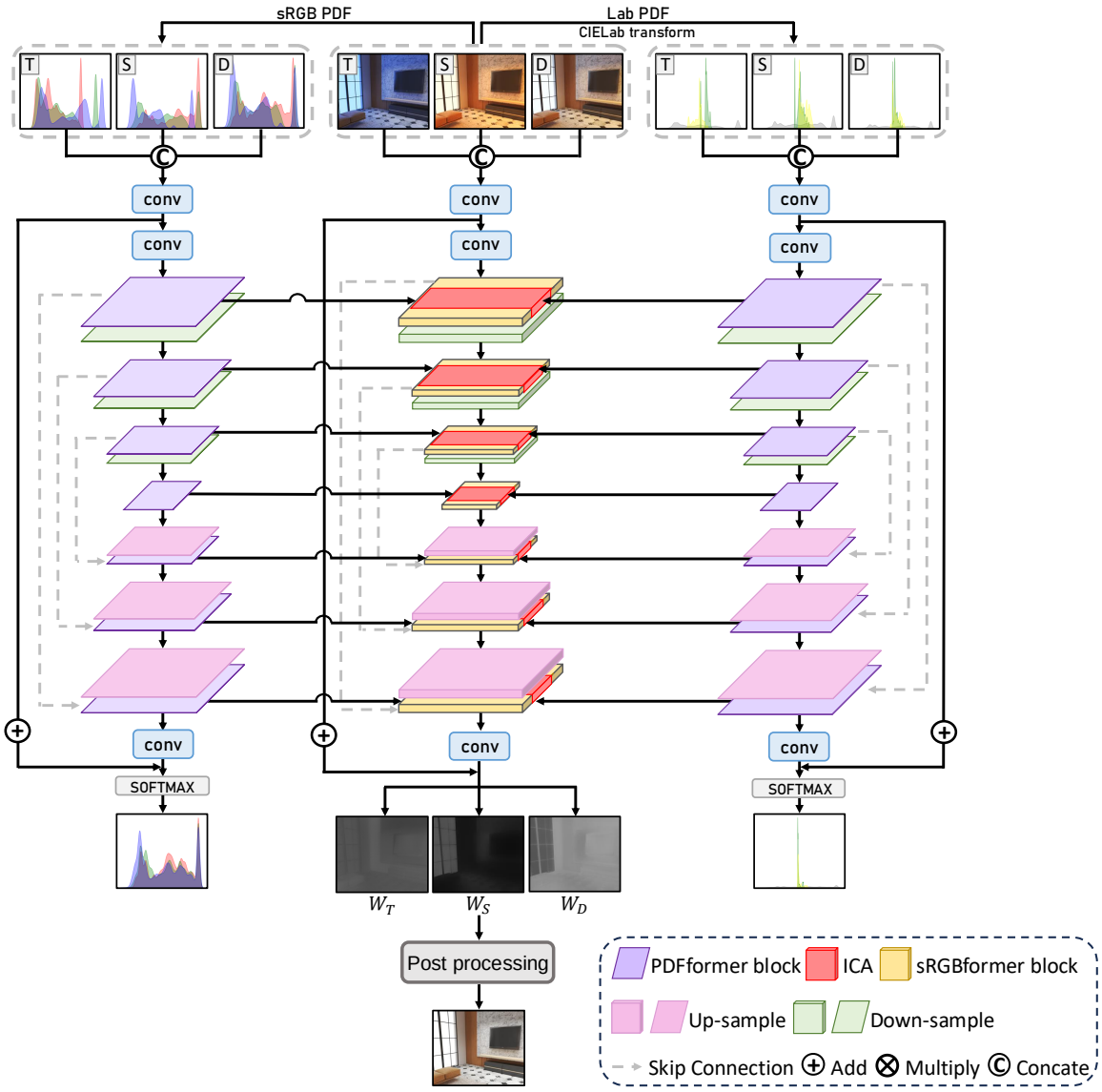

Figure 1. The extended ABC-Former, **ABC-FormerM**, addresses the multi-illuminant task. It processes multi-illuminant input images (e.g., **T**ungsten, **S**hade, and **D**aylight) and their corresponding histograms from the sRGB and CIELAB color spaces by concatenating them separately. These concatenated inputs are fed into the Auxiliary Models (PDFformers) and the Target Model (sRGBformer). ABC-FormerM outputs weighting maps ($\mathbf{W_T}$, $\mathbf{W_S}$, and $\mathbf{W_D}$) to blend the multi-illuminant images after post-processing, producing the final WB-corrected result.

Mean Angular Error (MAE), and ΔE 2000 [7]. Despite being a straightforward extension of our single-illuminant model, ABC-FormerM demonstrates competitive performance, particularly excelling in MSE and ΔE 2000. Figure 2 provides qualitative comparisons, showing that ABC-FormerM consistently achieves superior color correction in challenging multi-illuminant scenarios compared to existing methods.

3. More Experimental Results

3.1. Results on Multi-illuminant MIT-Adobe 5K Dataset

Multi-illuminant Methods. The MIT-Adobe 5K dataset [6] contains 5,000 raw images captured by professional and semi-professional photographers, covering diverse scenes, subjects, and color temperatures. We compare our ABC-FormerM with other multi-illuminant methods, including the standard Camera Auto White Bal-

Table 1. The WB results on the Mixed-illuminant dataset [3] are presented, using the mean, first quartile (Q1), second quartile (Q2), and third quartile (Q3) of the MSE, MAE, and ΔE 2000. The patch size is indicated by "p." To highlight the best results, colored values are used: Red for the best, Blue for the second-best, and Green for the third-best.

Method	MSE ↓				MAE ↓				ΔE 2000 ↓				Size
	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3	Mean	Q1	Q2	Q3	MB
KNN [2]	1226.57	680.65	1062.64	1573.89	5.81°	4.28°	5.76°	6.85°	12.00	9.37	11.56	13.61	21.8
Deep-WB [1]	1130.60	621.00	886.32	1274.72	**4.53°**	**3.55°**	**4.19°**	**5.21°**	10.93	**8.59**	**9.82**	11.96	16.7
WBFlow [9]	1105.38	672.03	962.54	1321.69	5.43°	3.86°	5.18°	6.31°	11.01	8.91	10.47	13.26	30.2
ABC-Former	941.06	**463.42**	**734.55**	1166.88	**4.95°**	3.91°	**4.68°**	5.71°	10.71	**8.53**	10.22	12.04	20.2
Mixed-WB [3]													
p = 64, WB = t,d,s	819.47	655.88	845.79	1000.82	5.43°	4.27°	4.89°	6.23°	10.61	9.42	10.72	11.81	5.09
p = 64, WB = t,f,d,c,s	938.02	757.29	961.55	1161.52	**4.67°**	**3.71°**	**4.14°**	**5.35°**	12.26	10.80	11.25	12.76	5.10
p = 128, WB = t,d,s	830.20	584.77	853.01	**992.56**	5.03°	3.93°	4.78°	5.90°	11.41	9.76	11.39	12.53	5.09
p = 128, WB = t,f,d,c,s	1089.69	846.21	1125.59	1279.39	5.64°	4.15°	5.09°	6.50°	13.75	11.45	12.58	15.59	5.10
Style WB [8]													
p = 64, WB = t,d,s	868.01	649.36	889.00	1026.98	5.73°	4.48°	5.42°	6.34°	12.11	10.42	12.12	13.36	61.0
p = 64, WB = t,f,d,c,s	1051.07	760.86	1024.00	1332.50	6.30°	4.43°	6.01°	7.69°	14.43	11.90	13.11	16.15	61.1
p = 128, WB = t,d,s	822.77	576.52	840.67	1025.26	5.11°	3.93°	4.85°	**5.51°**	11.65	10.63	11.86	13.02	61.2
p = 128, WB = t,f,d,c,s	834.28	625.95	8442.71	1005.59	5.71°	4.57°	5.54°	6.19°	11.79	9.84	12.19	13.00	61.3
ABC-FormerM													
p = 64, WB = t,d,s	**771.09**	528.17	831.16	996.09	5.58°	4.22°	5.66°	6.82°	**10.20**	8.74	10.60	**11.67**	20.2
p = 64, WB = t,f,d,c,s	**756.57**	**514.13**	**794.29**	**930.43**	5.21°	**3.79°**	5.04°	6.32°	10.21	8.83	10.42	11.78	20.2
p = 128, WB = t,d,s	773.80	528.12	830.45	1008.02	5.87°	4.53°	5.82°	6.93°	**10.06**	8.61	**10.12**	**11.61**	20.2
p = 128, WB = t,f,d,c,s	**750.72**	**526.05**	**791.05**	**951.49**	5.48°	4.08°	5.72°	6.69°	**9.85**	**8.52**	**10.11**	**11.26**	20.2

Figure 2. Qualitative comparisons with other WB methods on the Mixed-illuminant dataset [3], with the ΔE 2000 displayed at the bottom of each corrected image.

ance (AWB), Mixed-WB [3], and Style WB [8], as shown in Figures 4-7. To simulate multi-illuminant inputs, raw images were processed with various WB settings ($\{t, d, s\}$ or $\{t, f, d, c, s\}$) in Adobe Photoshop.

Single-illuminant Methods.

We also compare our ABC-Former (single-illuminant

version) with Camera AWB and other single-illuminant methods, including KNN [2], Deep-WB [1], and WBFlow [9], in Figures 8-11. The results demonstrate that our method effectively corrects color casts and achieves natural-looking WB, even in multi-illuminant scenarios.

3.2. Additional Results on Single-illuminant Datasets

To further evaluate WB correction, we present qualitative comparisons of Deep-WB [1], WBFlow [9], and our ABC-Former on images from the Rendered WB dataset Set1-Test and Set2 [2] and the Rendered Cube+ dataset [2, 4]. Results are shown in Figures 12-15 (Rendered WB dataset Set1-Test), Figures 16-19 (Rendered WB dataset Set2), and Figures 20-23 (Rendered Cube+ dataset). The comparisons highlight that ABC-Former consistently restores natural colors, proving its effectiveness in diverse color correction tasks.

3.3. Analysis of WB Performance vs. Model Size

Figure 3 compares the mean ΔE 2000 scores and model sizes of various WB models [1–3, 9, 10] on the Rendered Cube+ dataset [2, 4]. We use the ΔE 2000 metric [7] for its alignment with human visual perception and sensitivity to subtle color changes. The figure shows that ABC-Former outperforms competing models while maintaining a compact size, demonstrating its efficiency and effectiveness.

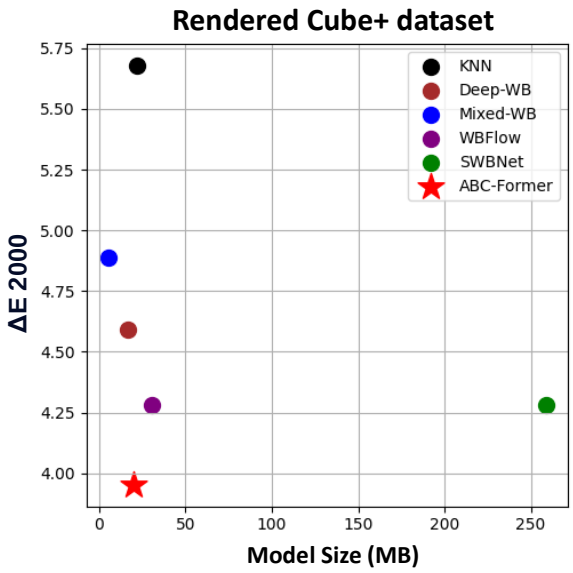

Figure 3. The model size vs. ΔE 2000 performance of WB models on the Rendered Cube+ dataset [2, 4].

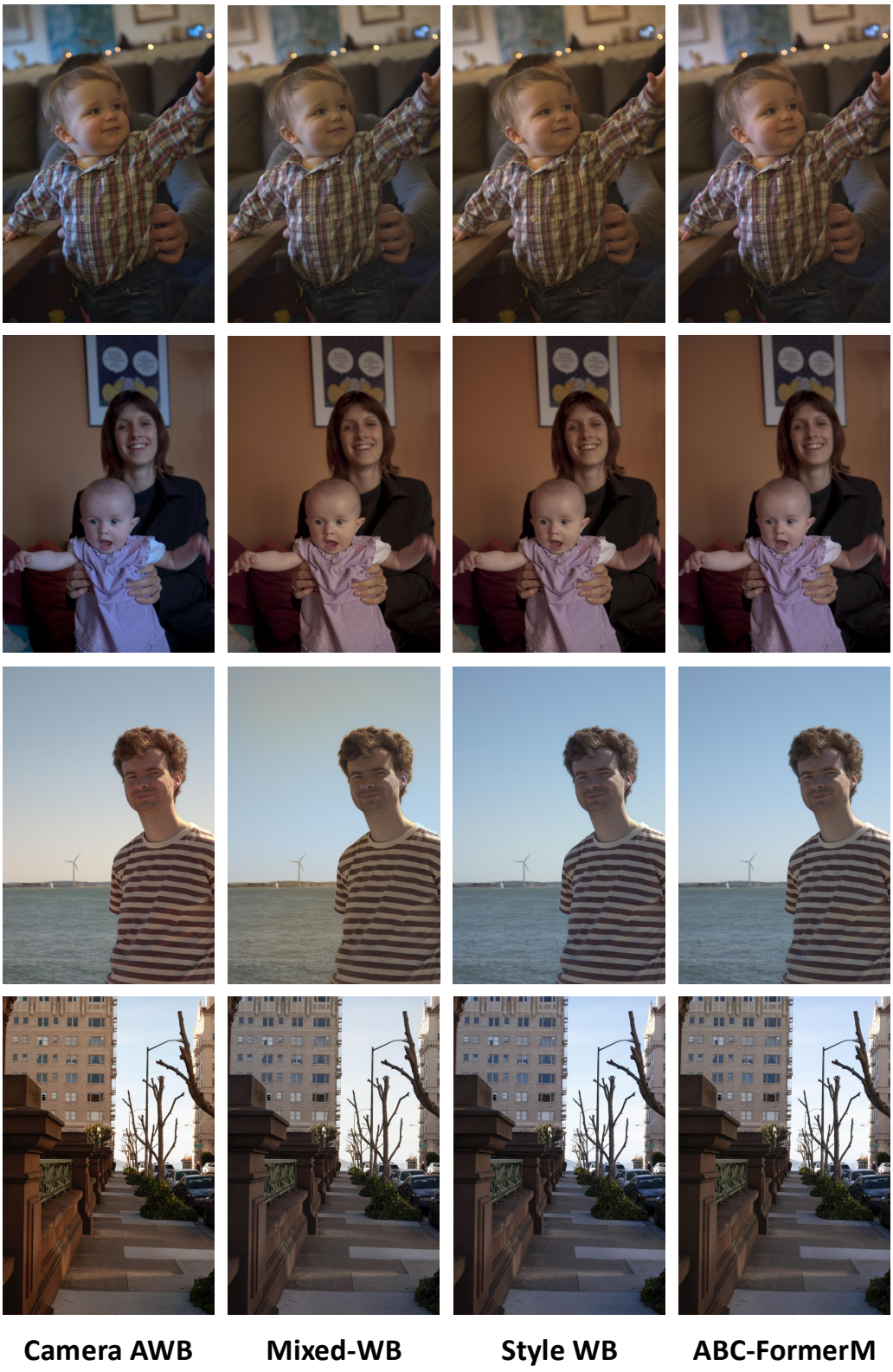

Figure 4. Qualitative comparisons with multi-input WB methods on the MIT-Adobe 5K dataset [6].

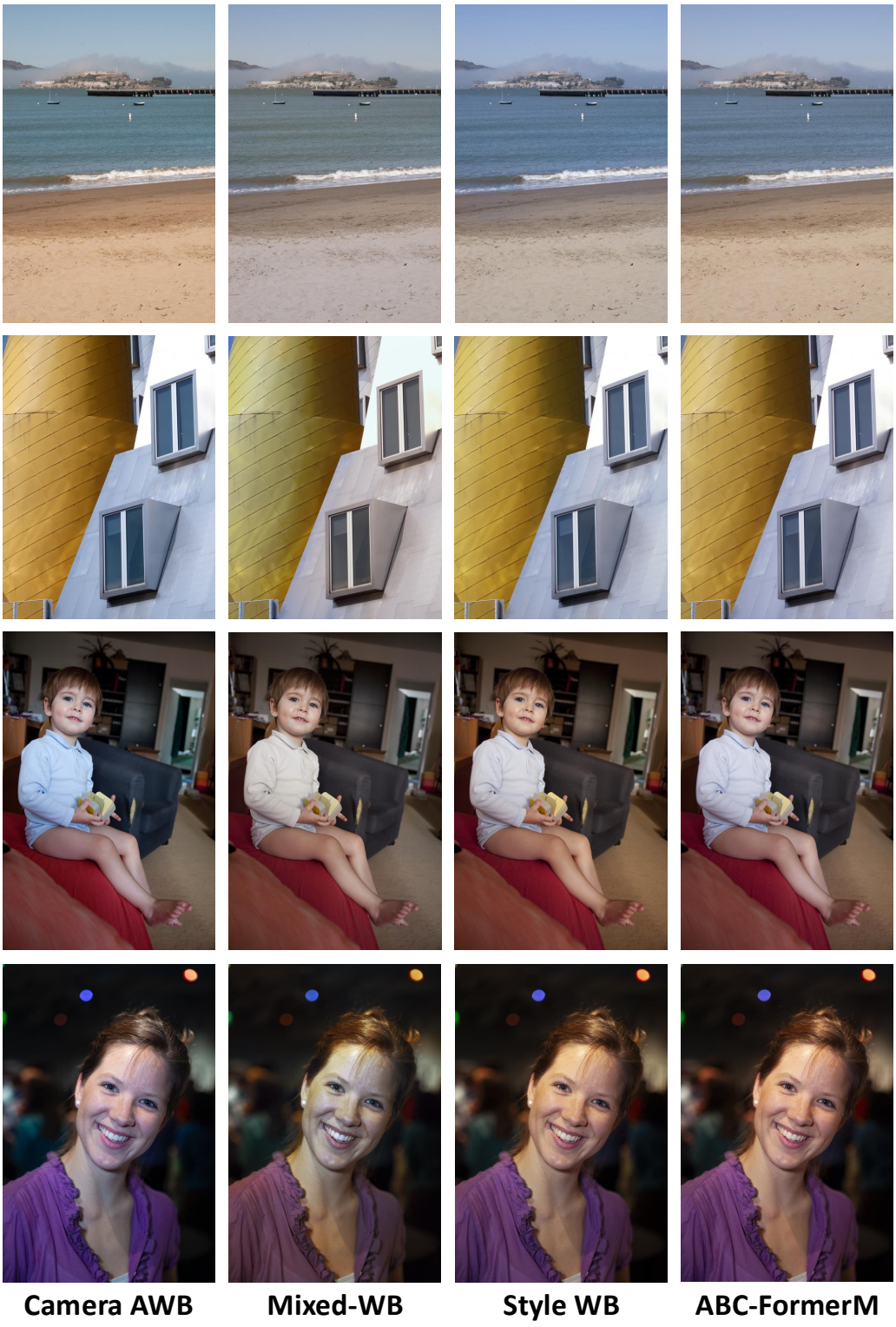

Figure 5. Qualitative comparisons with multi-input WB methods on the MIT-Adobe 5K dataset [6].

Figure 6. Qualitative comparisons with multi-input WB methods on the MIT-Adobe 5K dataset [6].

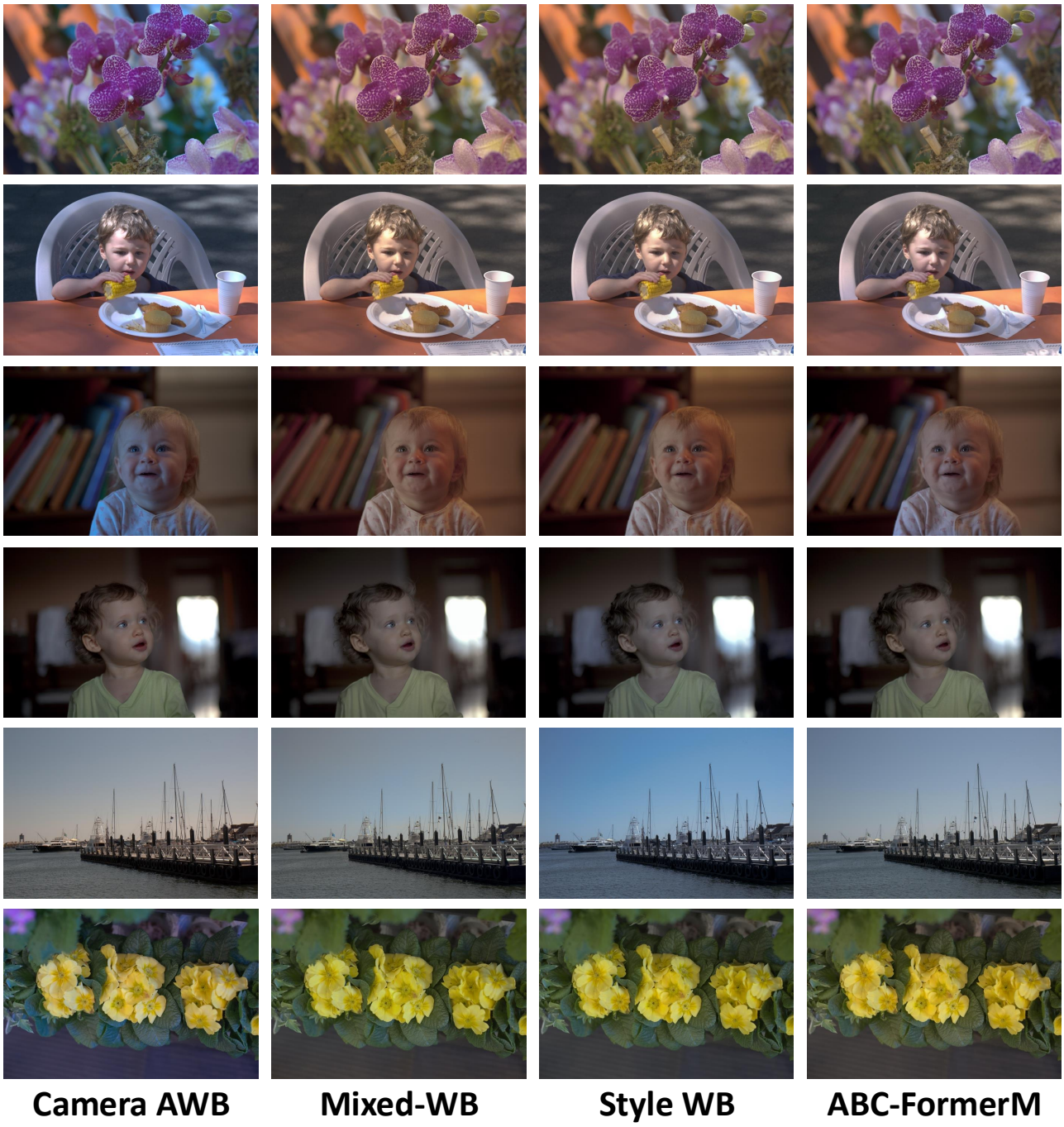

Figure 7. Qualitative comparisons with multi-input WB methods on the MIT-Adobe 5K dataset [6].

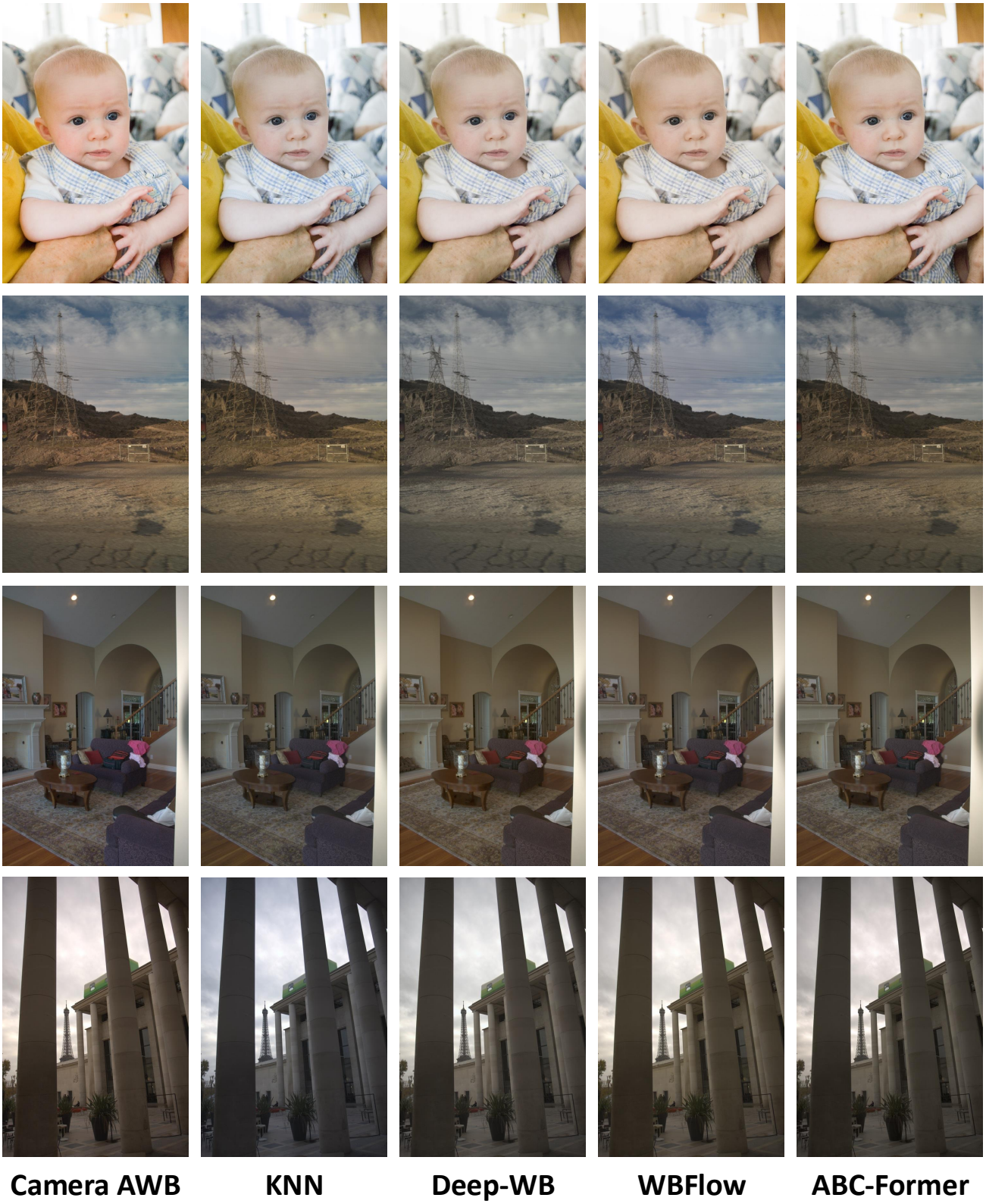

Figure 8. Qualitative comparisons with single-input WB methods on the MIT-Adobe 5K dataset [6].

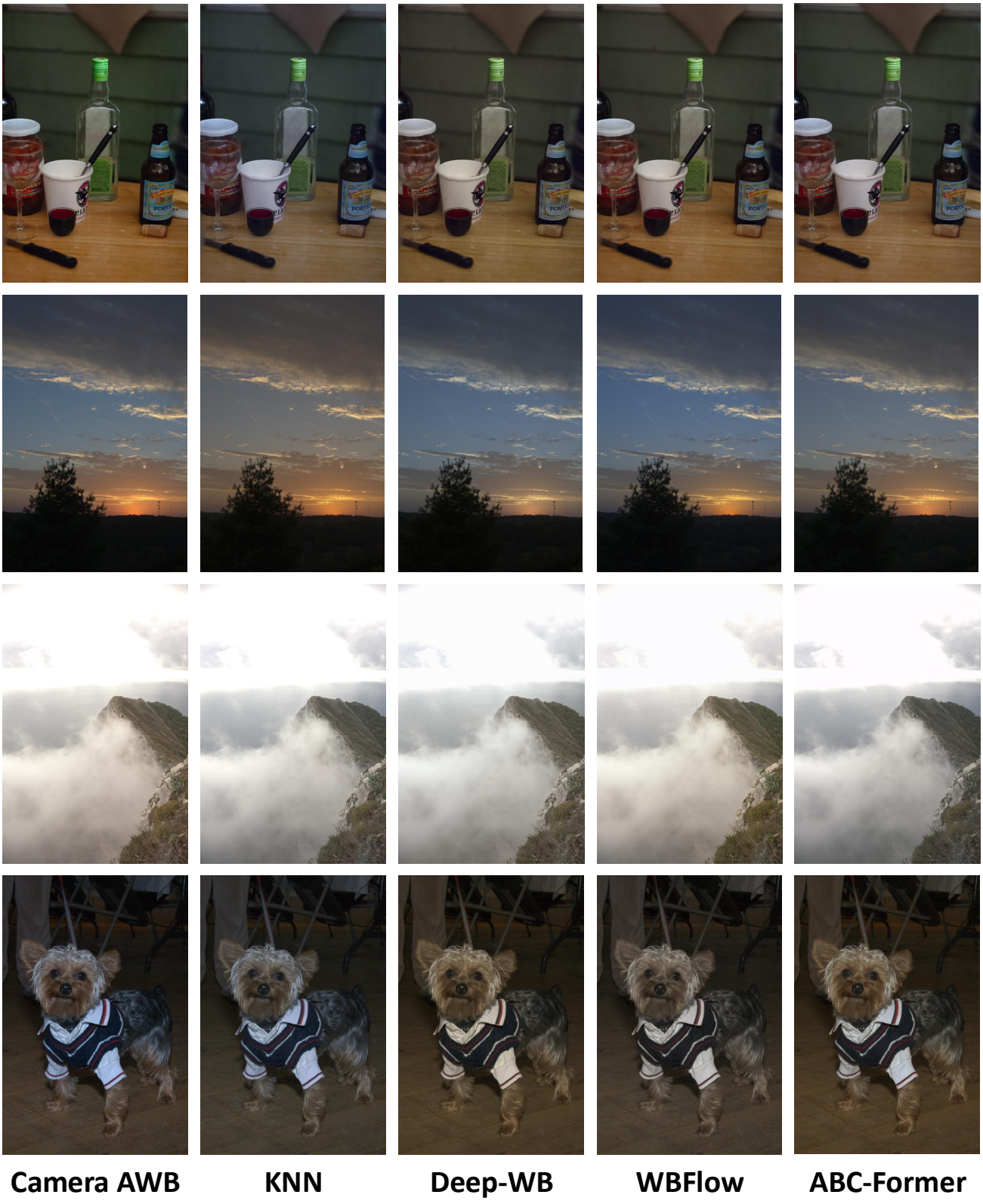

Figure 9. Qualitative comparisons with single-input WB methods on the MIT-Adobe 5K dataset [6].

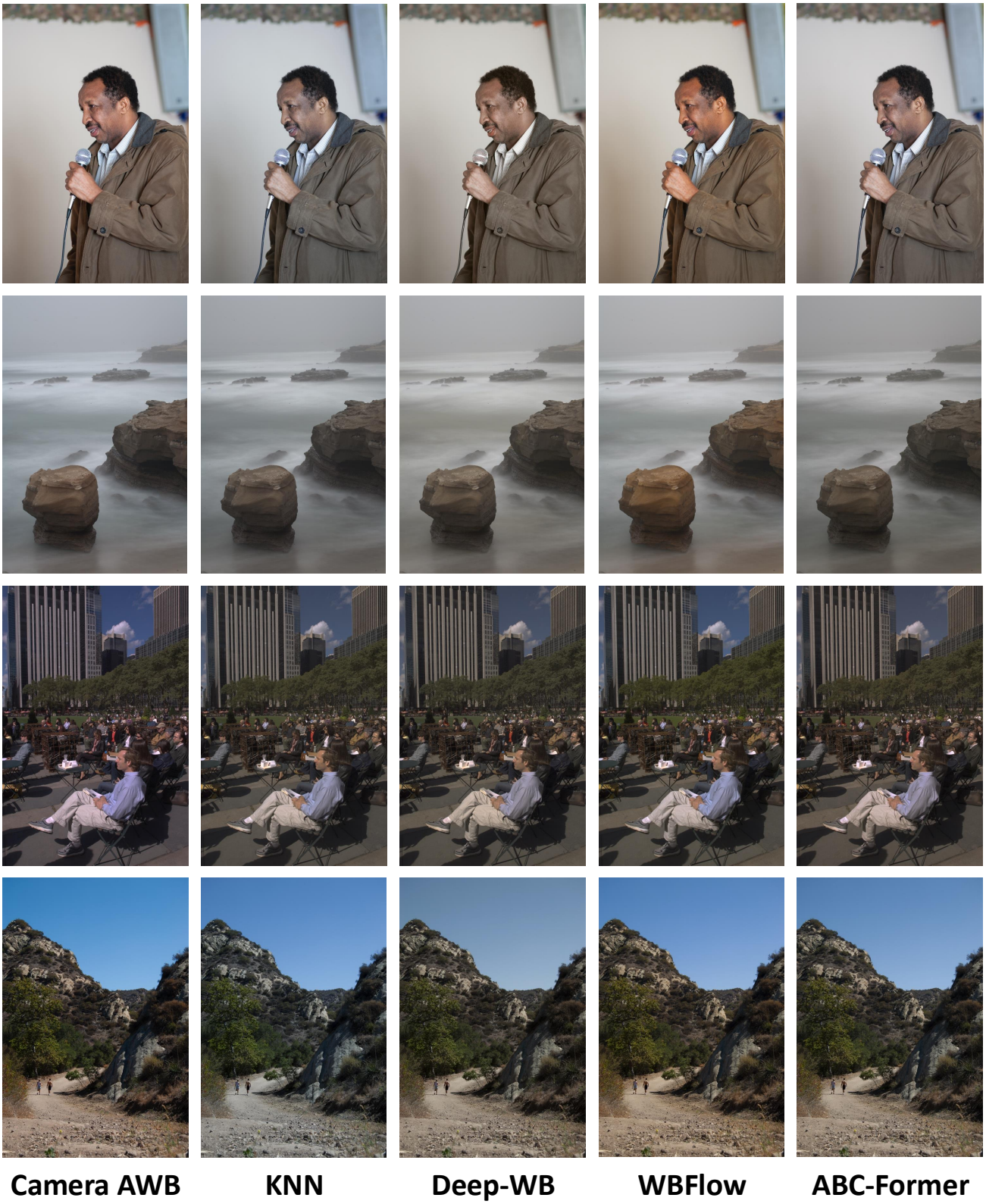

Figure 10. Qualitative comparisons with single-input WB methods on the MIT-Adobe 5K dataset [6].

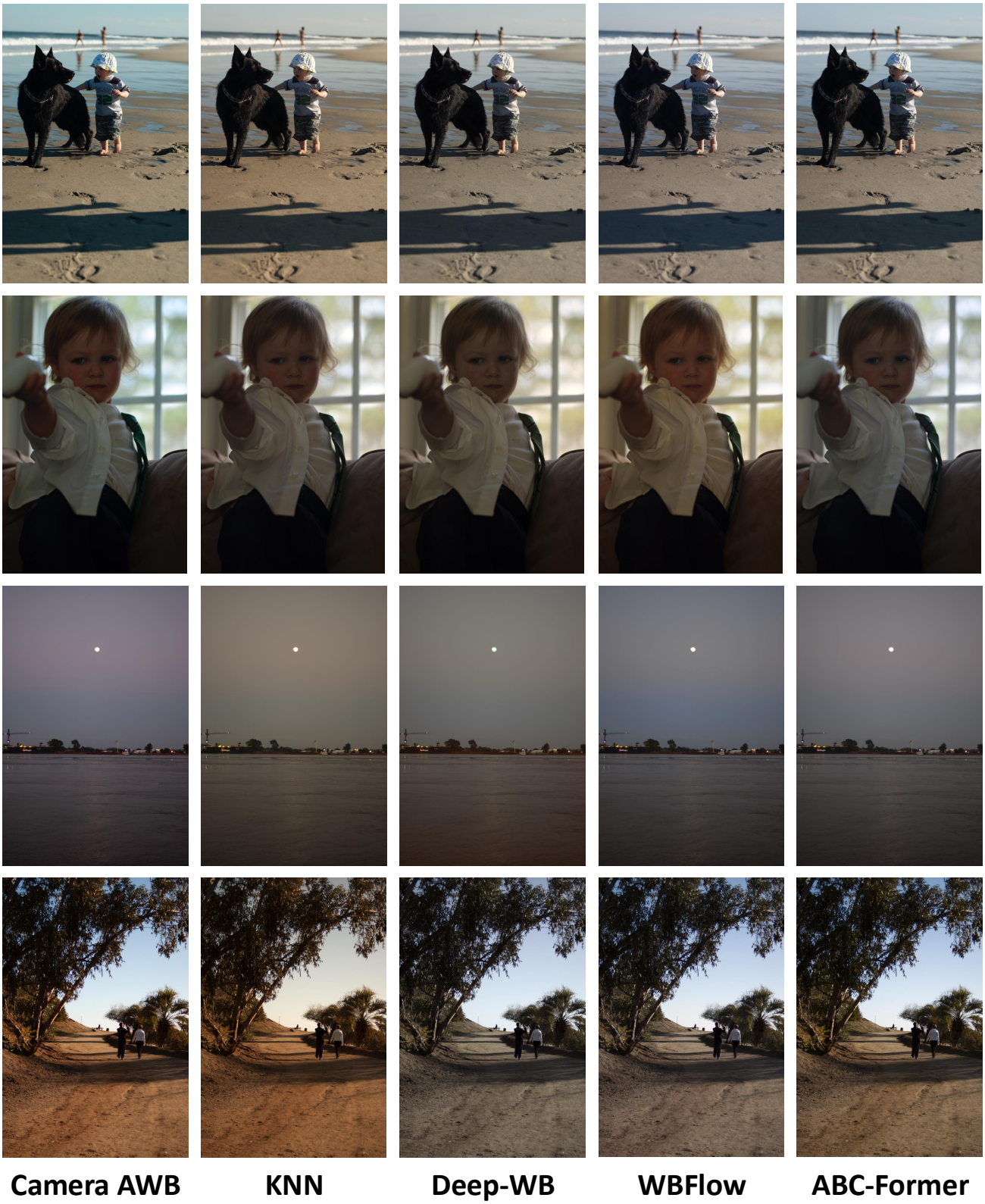

Figure 11. Qualitative comparisons with single-input WB methods on the MIT-Adobe 5K dataset [6].

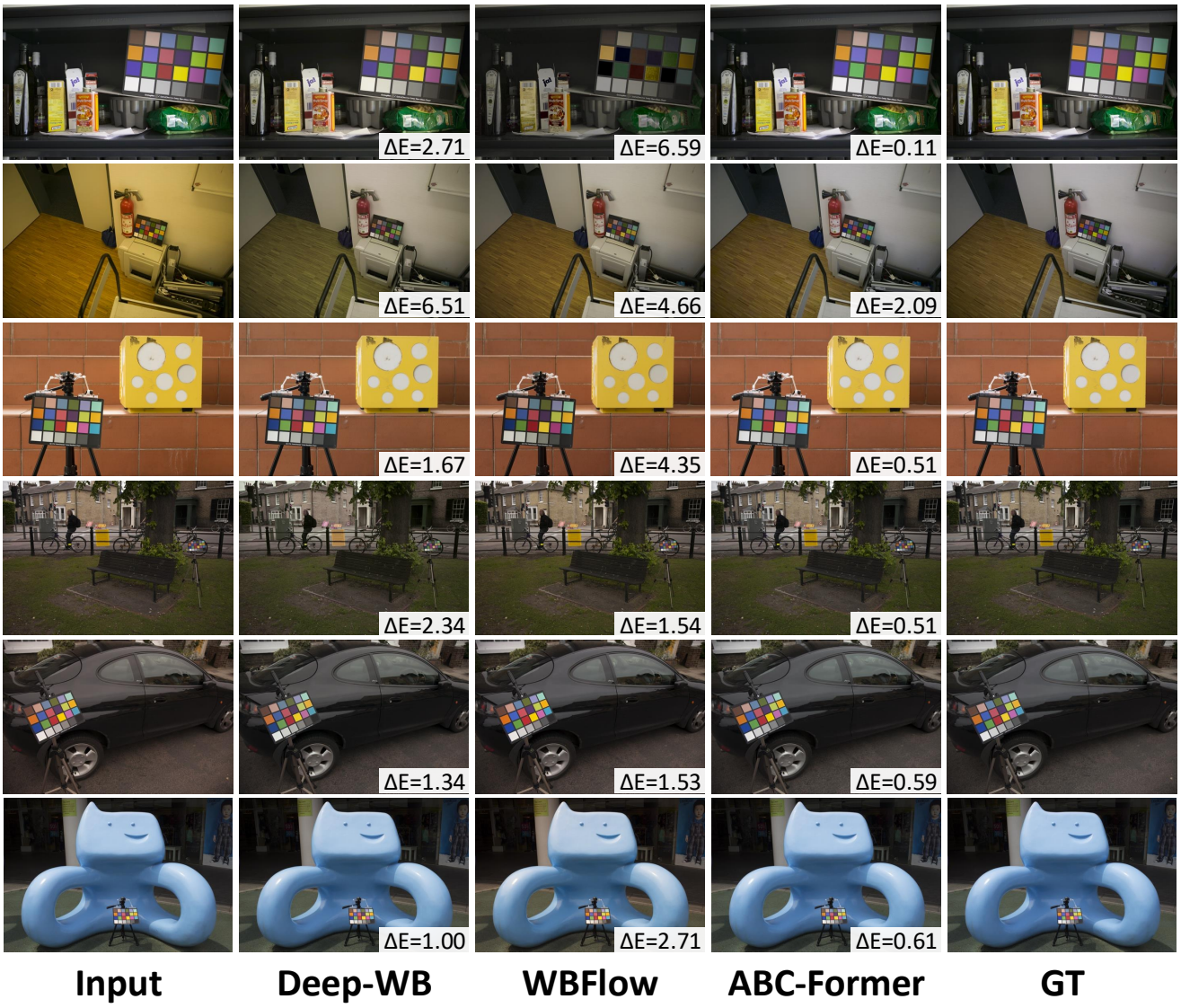

Figure 12. Qualitative comparisons of WB methods on the Rendered WB dataset Set1-Test [2], with ΔE 2000 shown in the bottom-right corner of each image.

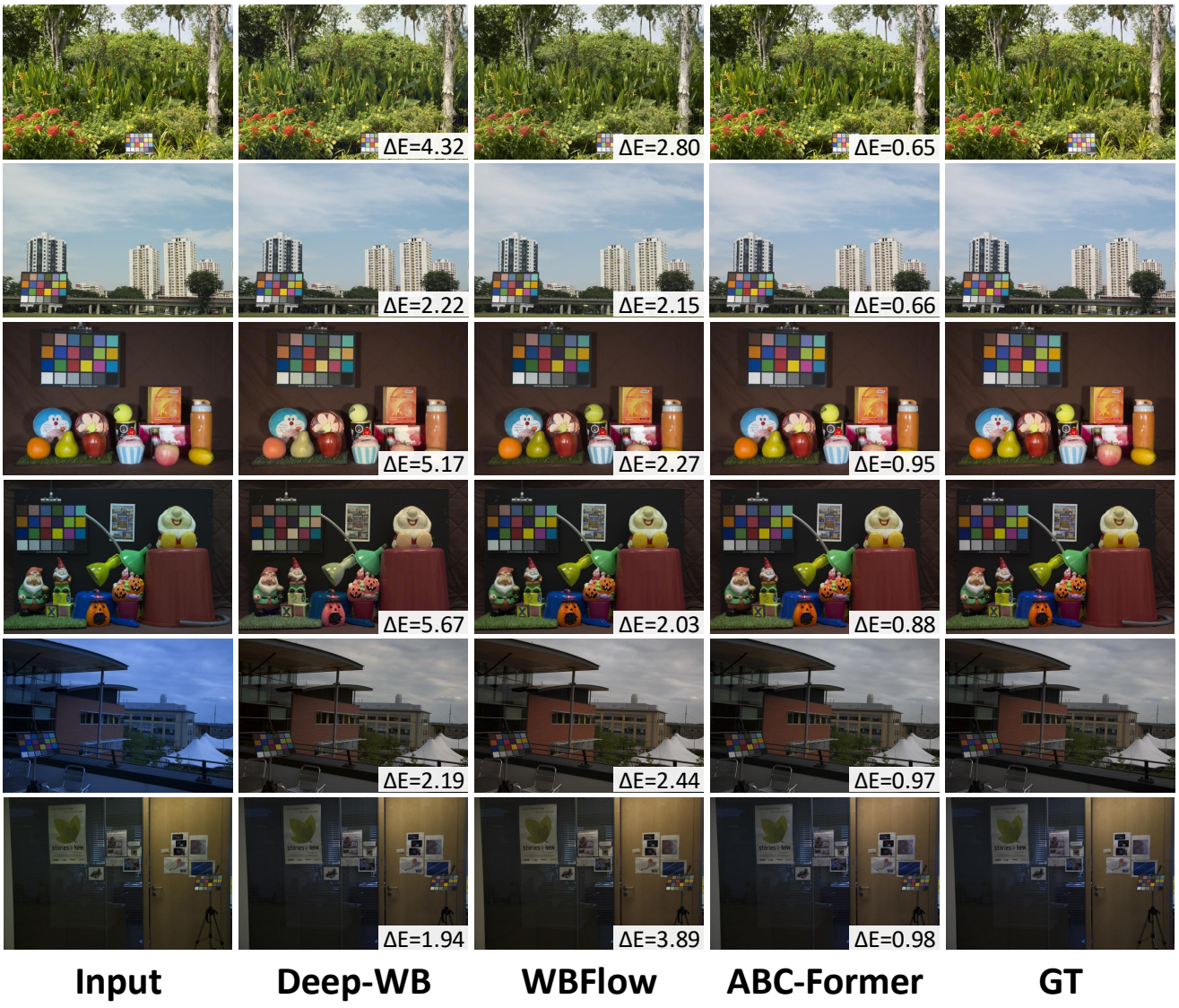

Figure 13. Qualitative comparisons of WB methods on the Rendered WB dataset Set1-Test [2], with ΔE 2000 shown in the bottom-right corner of each image.

Figure 14. Qualitative comparisons of WB methods on the Rendered WB dataset Set1-Test [2], with ΔE 2000 shown in the bottom-right corner of each image.

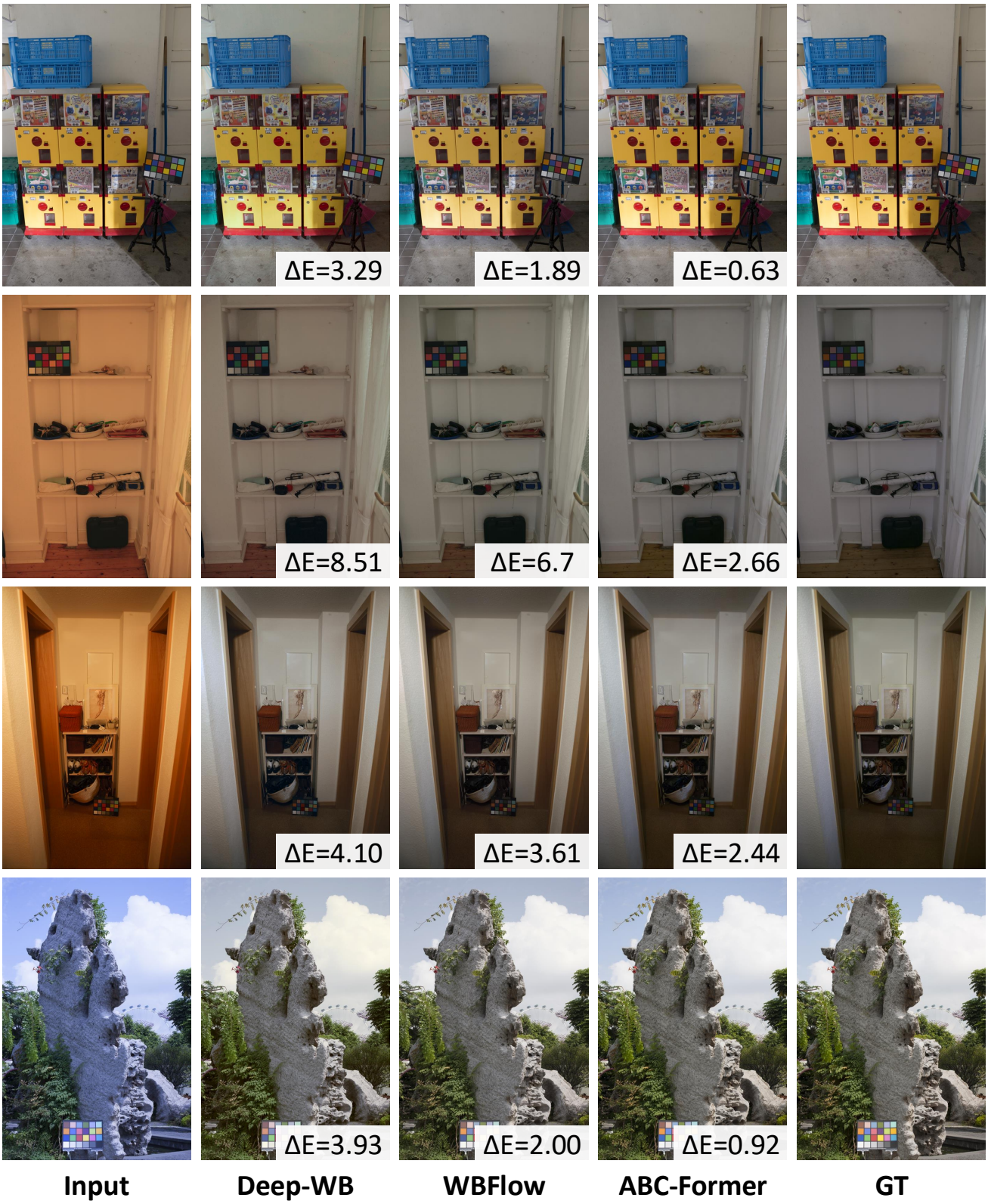

Figure 15. Qualitative comparisons of WB methods on the Rendered WB dataset Set1-Test [2], with ΔE 2000 shown in the bottom-right corner of each image.

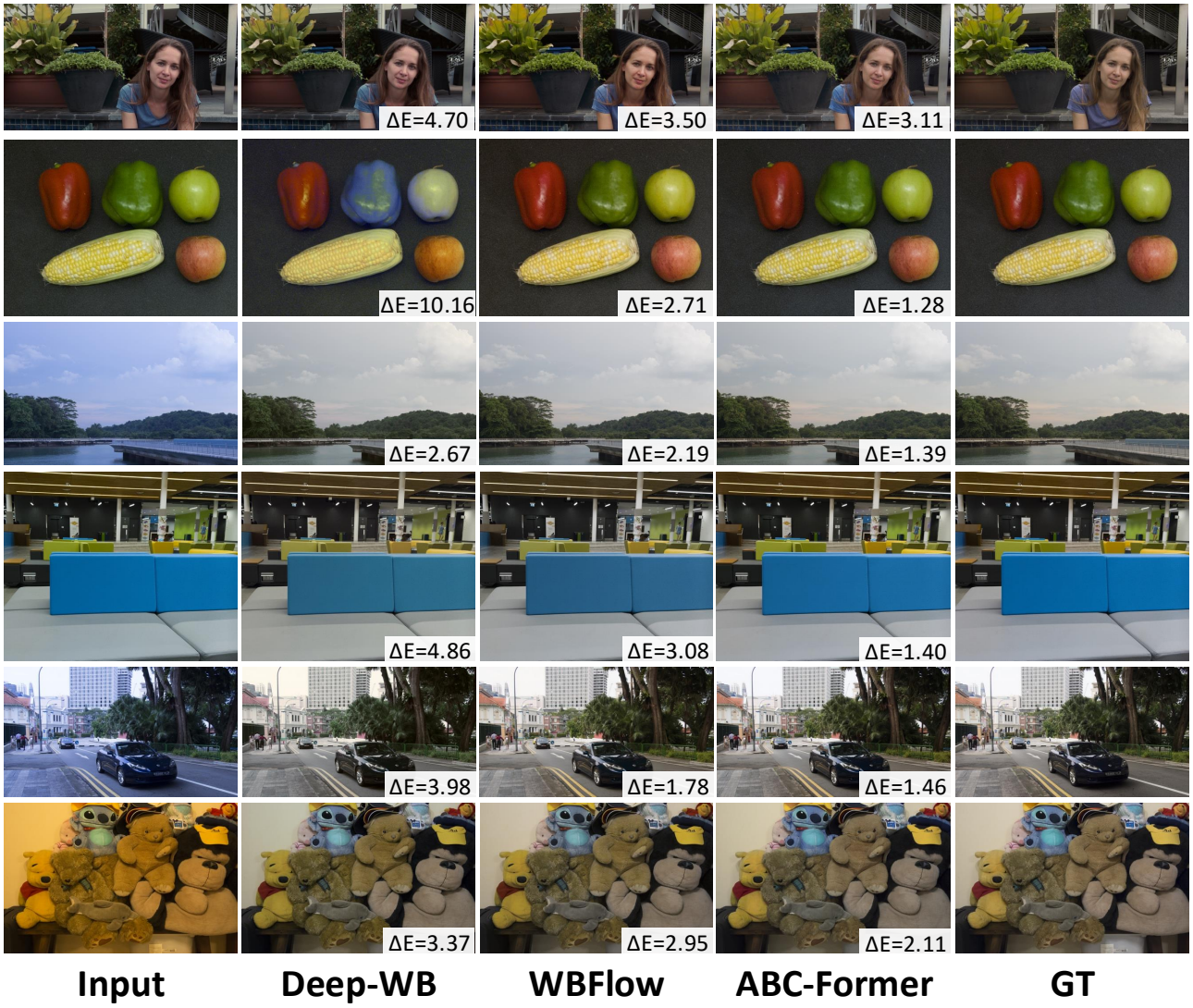

Figure 16. Qualitative comparisons of WB methods on the Rendered WB dataset Set2 [2], with ΔE 2000 shown in the bottom-right corner of each image.

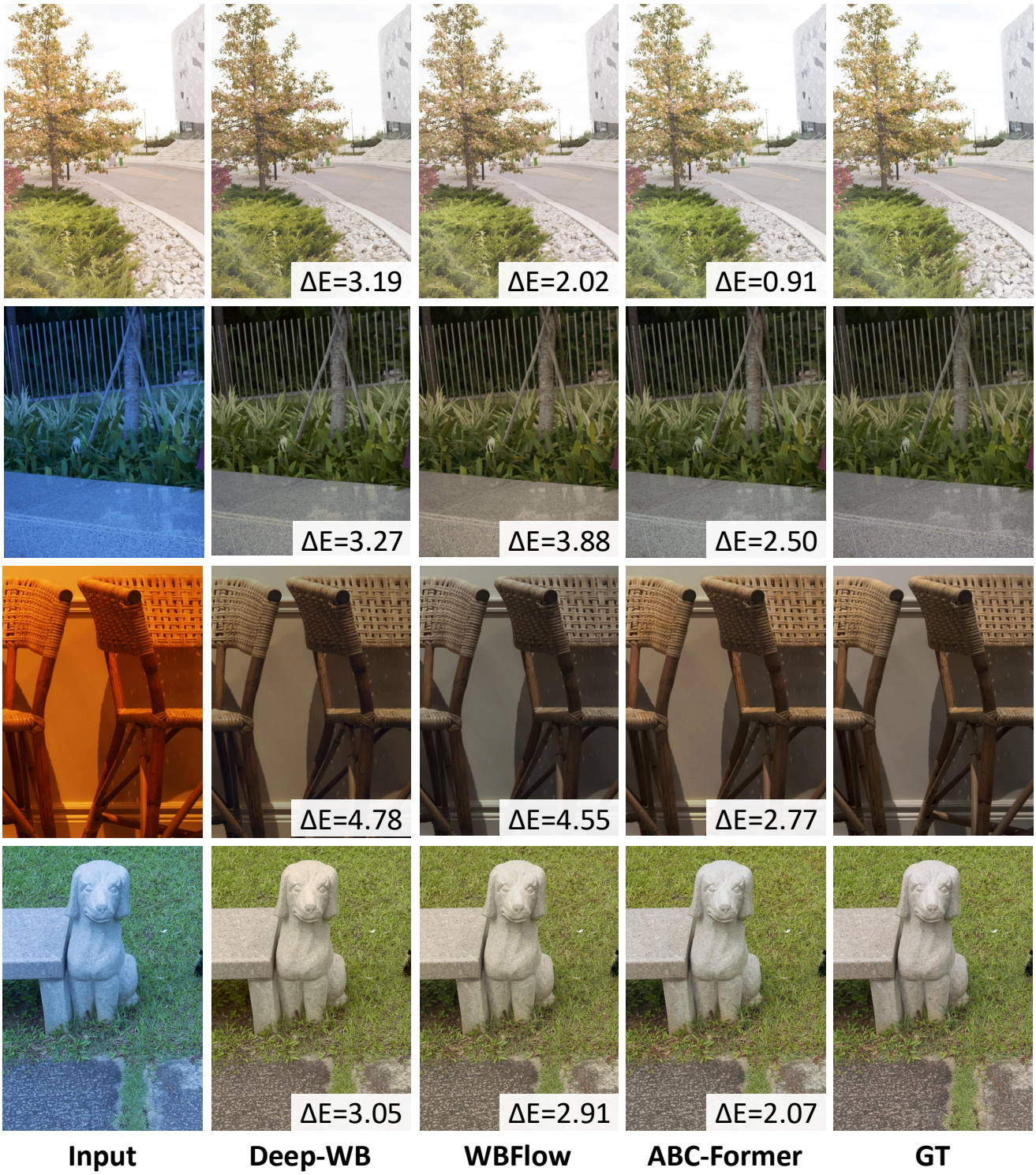

Figure 17. Qualitative comparisons of WB methods on the Rendered WB dataset Set2 [2], with ΔE 2000 shown in the bottom-right corner of each image.

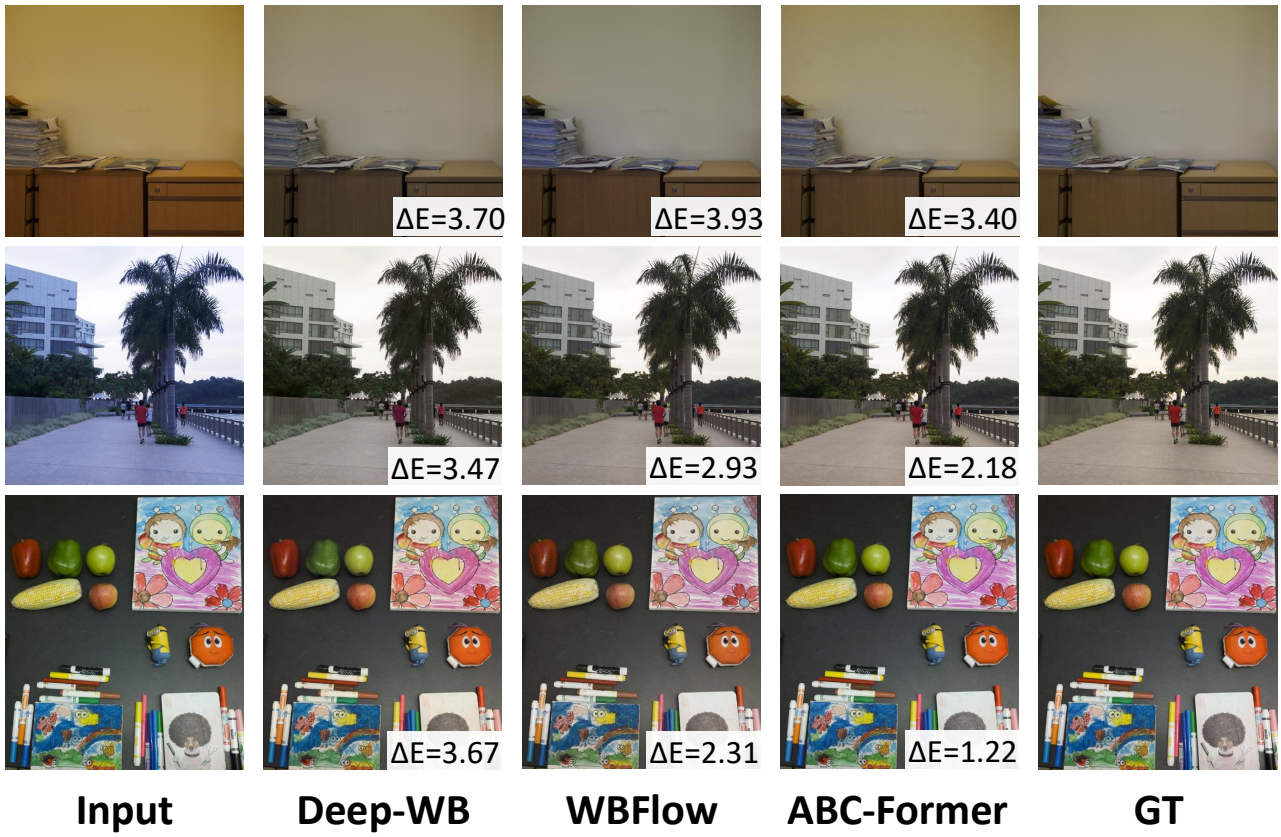

Figure 18. Qualitative comparisons of WB methods on the Rendered WB dataset Set2 [2], with ΔE 2000 shown in the bottom-right corner of each image.

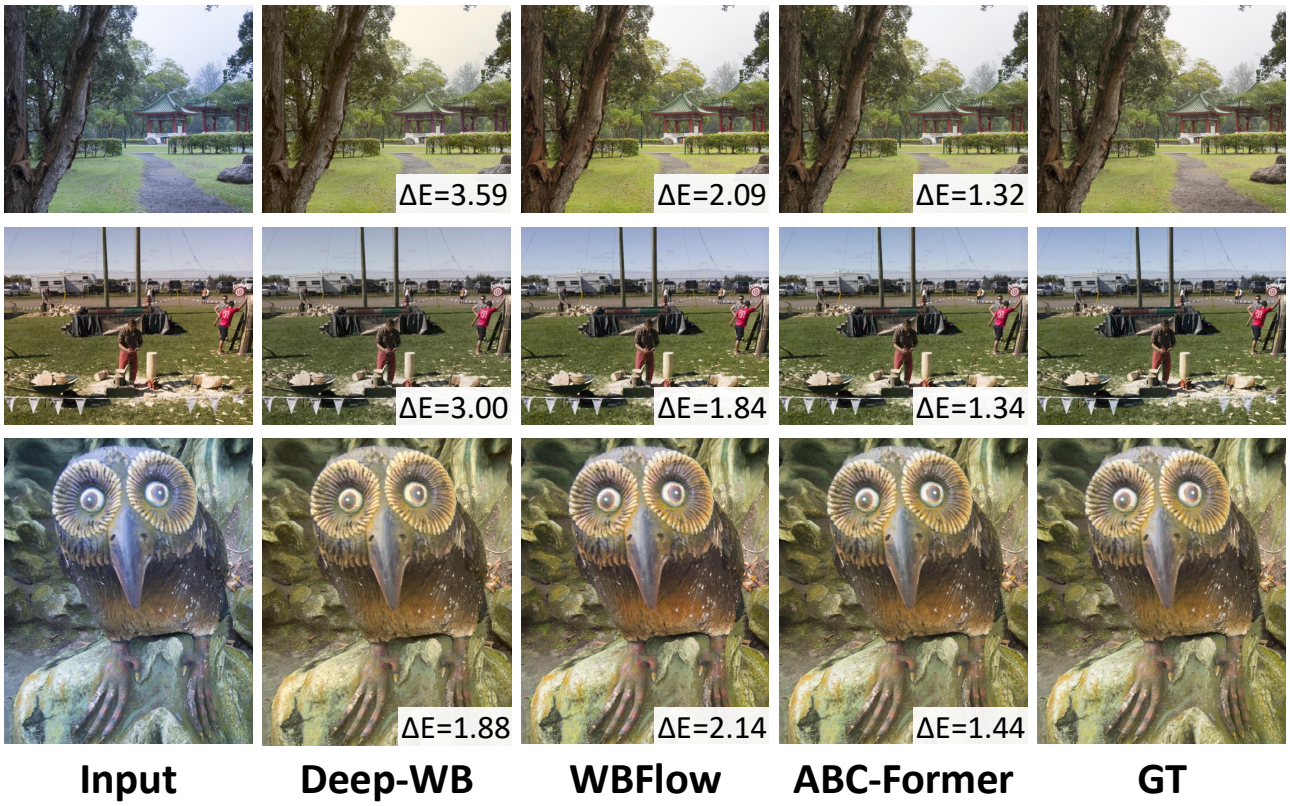

Figure 19. Qualitative comparisons of WB methods on the Rendered WB dataset Set2 [2], with ΔE 2000 shown in the bottom-right corner of each image.

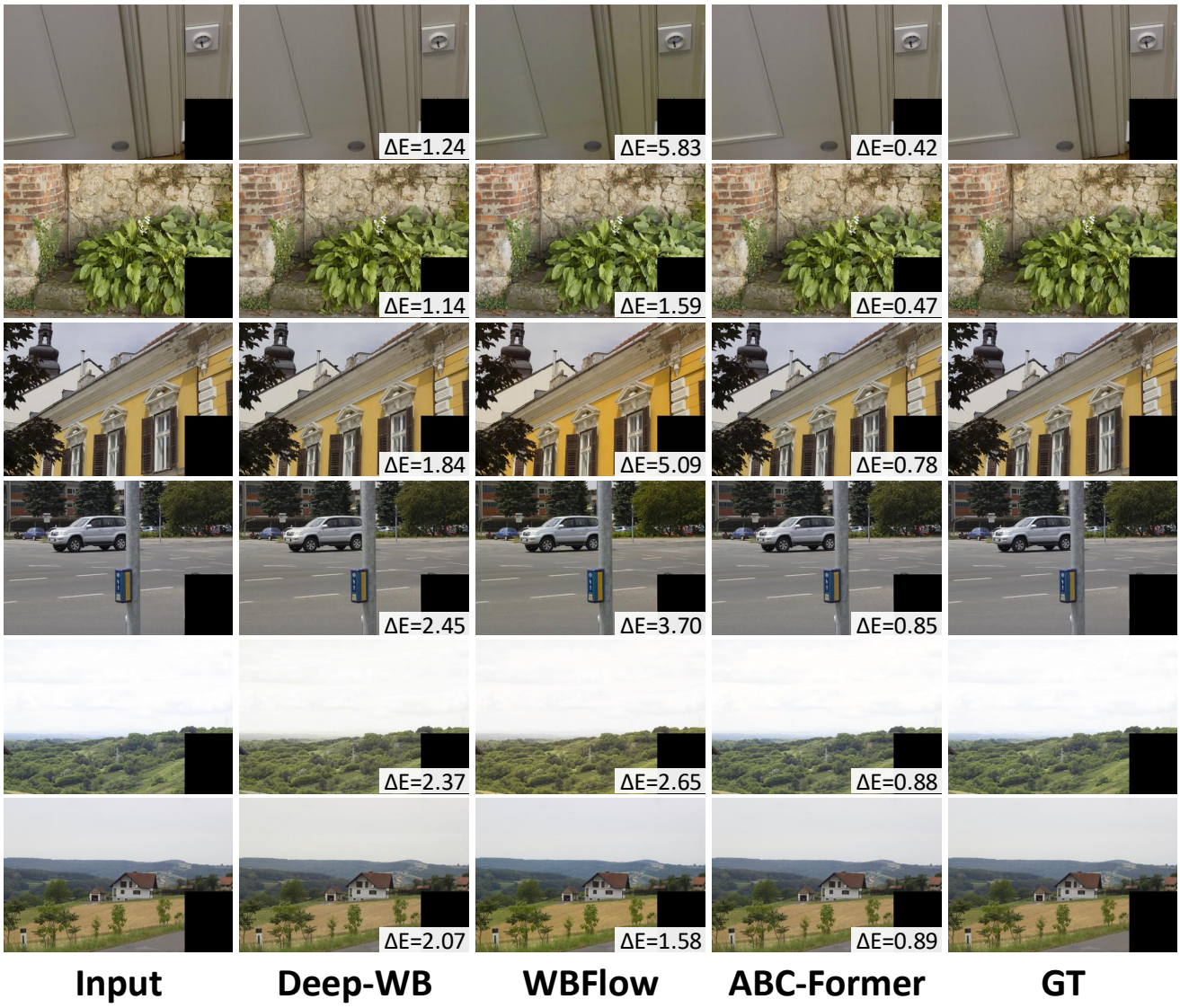

Figure 20. Qualitative comparisons of WB methods on the Rendered Cube+ dataset [2, 4], with ΔE 2000 shown in the bottom-right corner of each image.

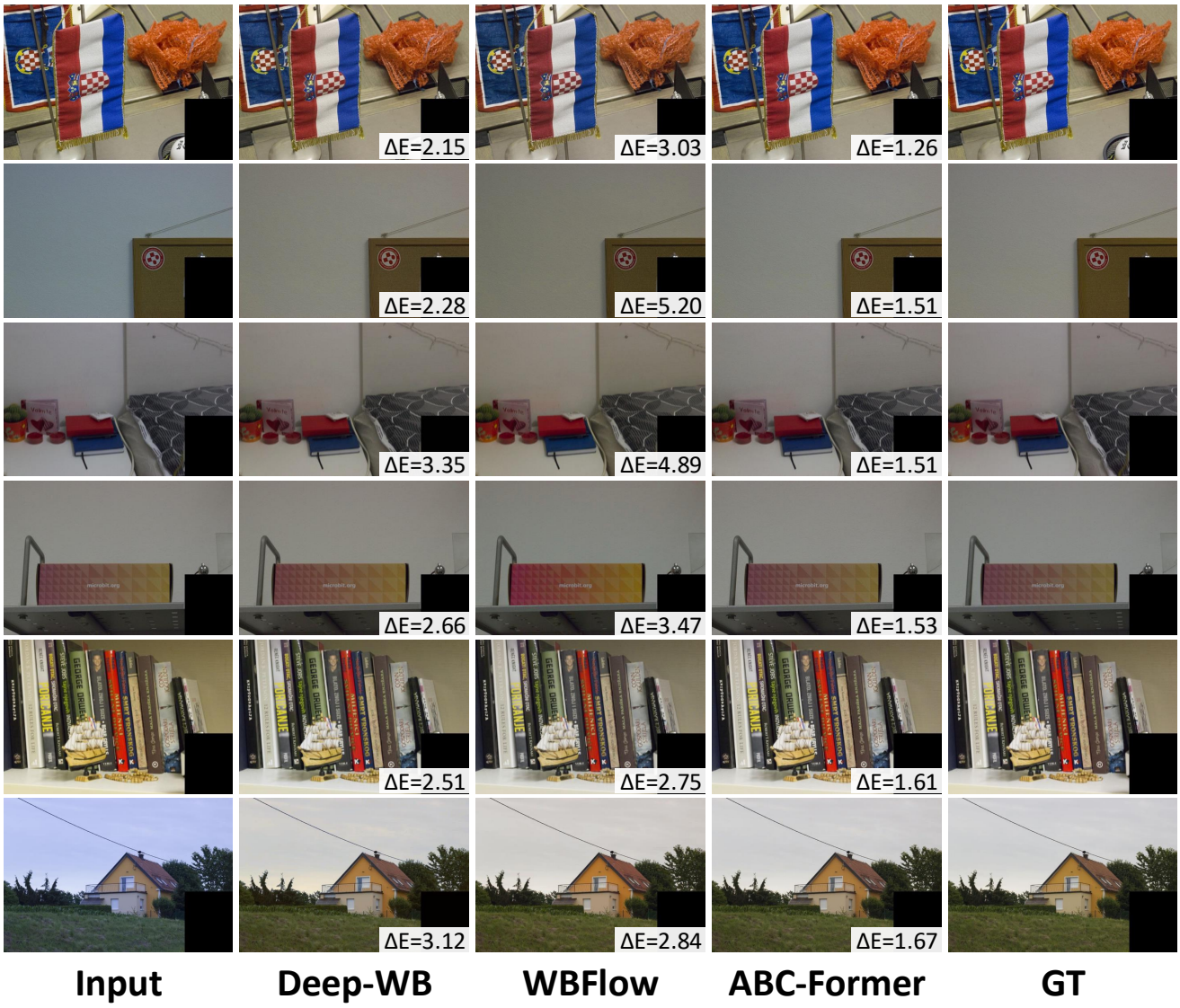

Figure 21. Qualitative comparisons of WB methods on the Rendered Cube+ dataset [2, 4], with ΔE 2000 shown in the bottom-right corner of each image.

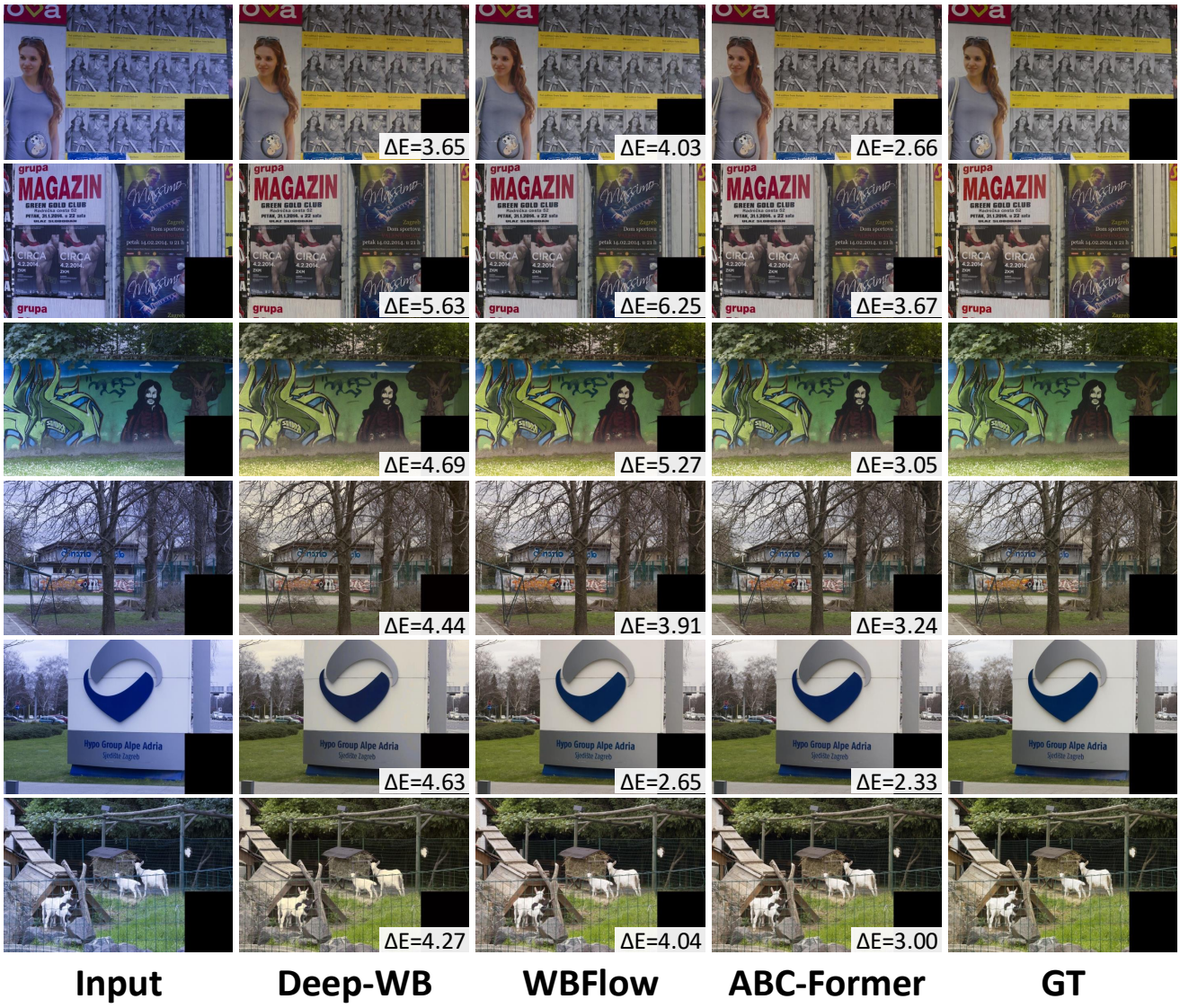

Figure 22. Qualitative comparisons of WB methods on the Rendered Cube+ dataset [2, 4], with ΔE 2000 shown in the bottom-right corner of each image.

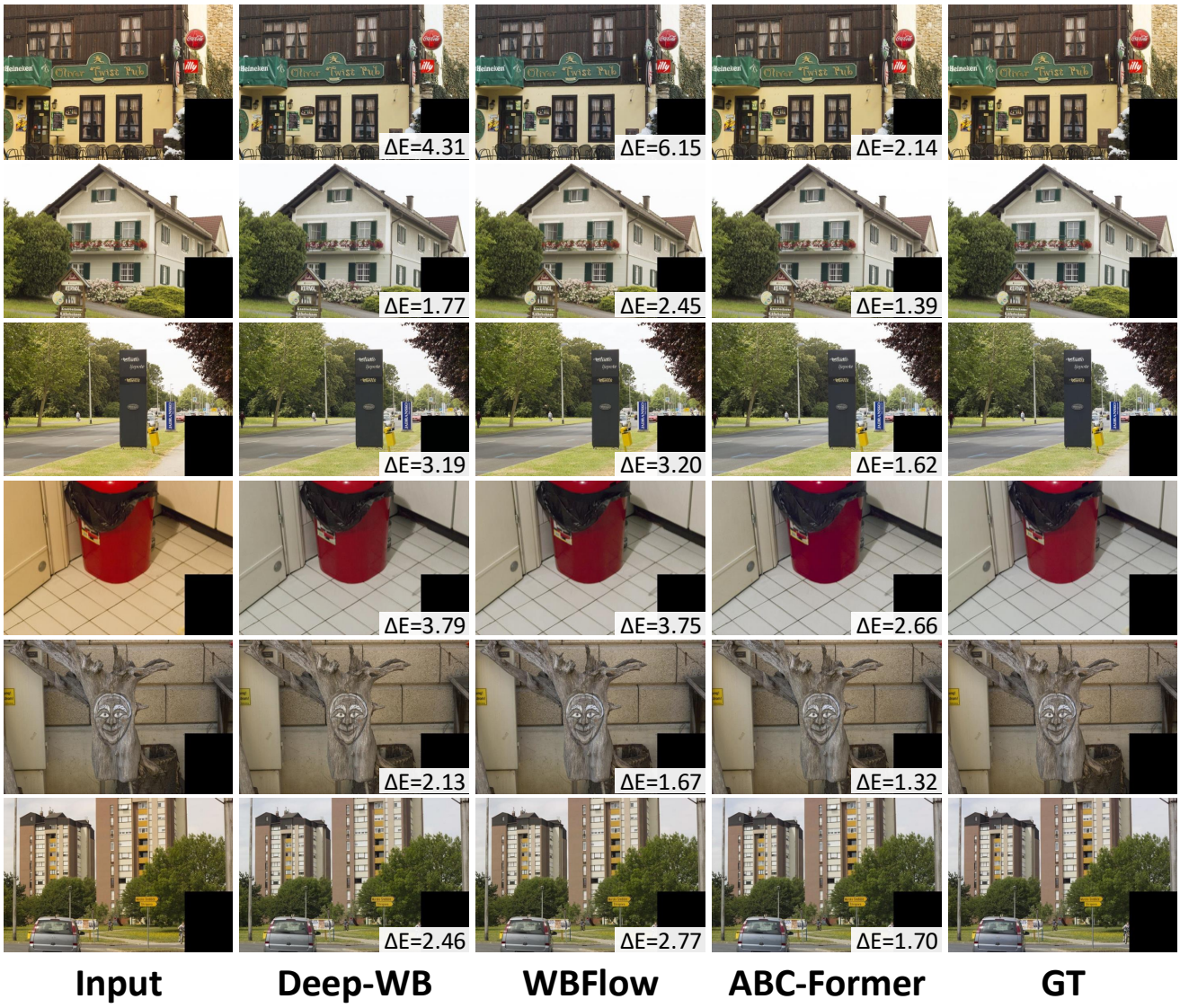

Figure 23. Qualitative comparisons with WB methods on the Rendered Cube+ dataset [2, 4], with the ΔE 2000 indicated in the bottom-right corner of each image.

References

[1] 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. 1, 3, 4

[2] 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. 1, 3, 4, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24

[3] 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. 1, 3, 4

[4] Nikola Banić, Karlo Koščević, and Sven Lončarić. Unsupervised learning for color constancy. *arXiv preprint arXiv:1712.00436*, 2017. 1, 4, 21, 22, 23, 24

[5] Jonathan T Barron and Ben Poole. The fast bilateral solver. In *European conference on computer vision*, pages 617–632. Springer, 2016. 1

[6] 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*, pages 97–104. IEEE, 2011. 1, 2, 5, 6, 7, 8, 9, 10, 11, 12

[7] Sharma Gaurav. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *COLOR research and application*, 30 (1):21–30, 2005. 2, 4

[8] Furkan Kınlı, Doğa Yılmaz, Barış Özcan, and Furkan Kıraç. Modeling the lighting in scenes as style for auto white-balance correction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4903–4913, 2023. 1, 3

[9] 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. 1, 3, 4

[10] 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. 4