Supplementary for Deep Image Harmonization with Globally Guided Feature Transformation and Relation Distillation

Li Niu[1*], Linfeng Tan[1], Xinhao Tao[1], Junyan Cao[1], Fengjun Guo[2], Teng Long[2], Liqing Zhang[1]
[1] Department of Computer Science and Engineering, MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[2] INTSIG
{ustcnewly,tanlinfeng,taoxinhao,Joy_C1}@sjtu.edu.cn, {fengjun_guo,mike_long}@intsig.net, zhang-lq@cs.sjtu.edu.cn

In this document, we provide additional materials to support our main paper. In Section 1, we provide more details of constructing ccHarmony dataset. In Section 2, we show more visual comparison results with baselines. In Section 3, we evaluate different methods on real composite images. In Section 4, we provide visualization results of ablation studies. In Section 5, we discuss the limitation of our method.

1. More Details of ccHarmony Dataset

As introduced in the main paper, the existing harmonization datasets [3, 7, 9, 5] may not faithfully reflect natural illumination variation. The Hday2night subdataset in iHarmony4 [3] captures a group of images for the same scene under different illumination conditions, which can reflect natural illumination variation. Nevertheless, such data collection is extremely expensive. Therefore, we explore a novel way to construct harmonization dataset ccHarmony to approximate natural illumination variation. When constructing our ccHarmony dataset, we collect real images with color checker, segment proper foregrounds, and perform color transfer for the foregrounds, yielding synthetic composite images. Next, we will introduce the above three steps: real image selection in Section 1.1, foreground segmentation in Section 1.2, and foreground color transfer in Section 1.3.

1.1. Real Image Selection

We first collect images with color checker (see Figure 1(a)) from NUS dataset [1] and Gehler dataset [4], in which each image is captured with a color checker placed in the scene to record illumination information. Then, we perform the following filtering steps. 1) We notice that these two datasets contain images capturing the same scene with similar camera viewpoints, so we perform near-duplicate removal to remove the images with duplicated content. 2) We observe that in some images, the color checker cannot represent the global illumination information of the whole image, for example, the color checker is placed in the shadow area. Therefore, we remove those images with misleading color checker. 3) Another issue is that the color checker should not be included in the final image harmonization dataset, because the color check may provide shortcut for the harmonization network. Therefore, we discard the images in which the color checker is placed near the image center, and crop the remaining images to obtain the possibly largest region without color checker (see Figure 1(b)).

After the above filtering steps, we have 350 real images.

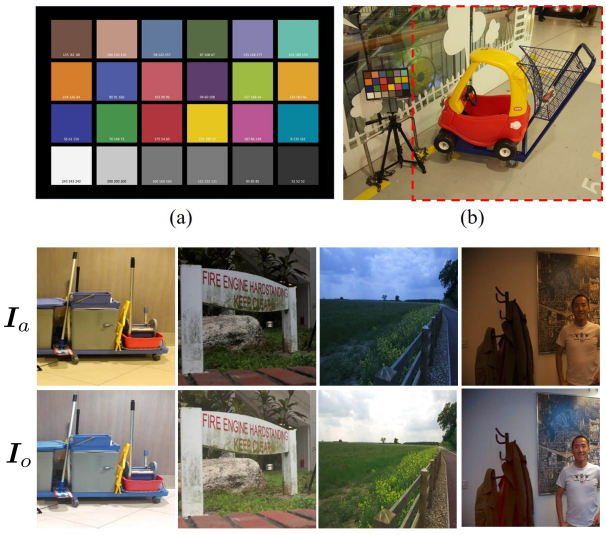

Figure 1. (a) Standard patch colors of a 24-patch Macbeth color checker. (b) An image captured with a color checker placed in the scene. The red dashed box indicates the cropped image without color checker, which is used in our dataset. (c) Examples of real images I_a and their counterparts I_o in standard illumination condition.

*Corresponding author.

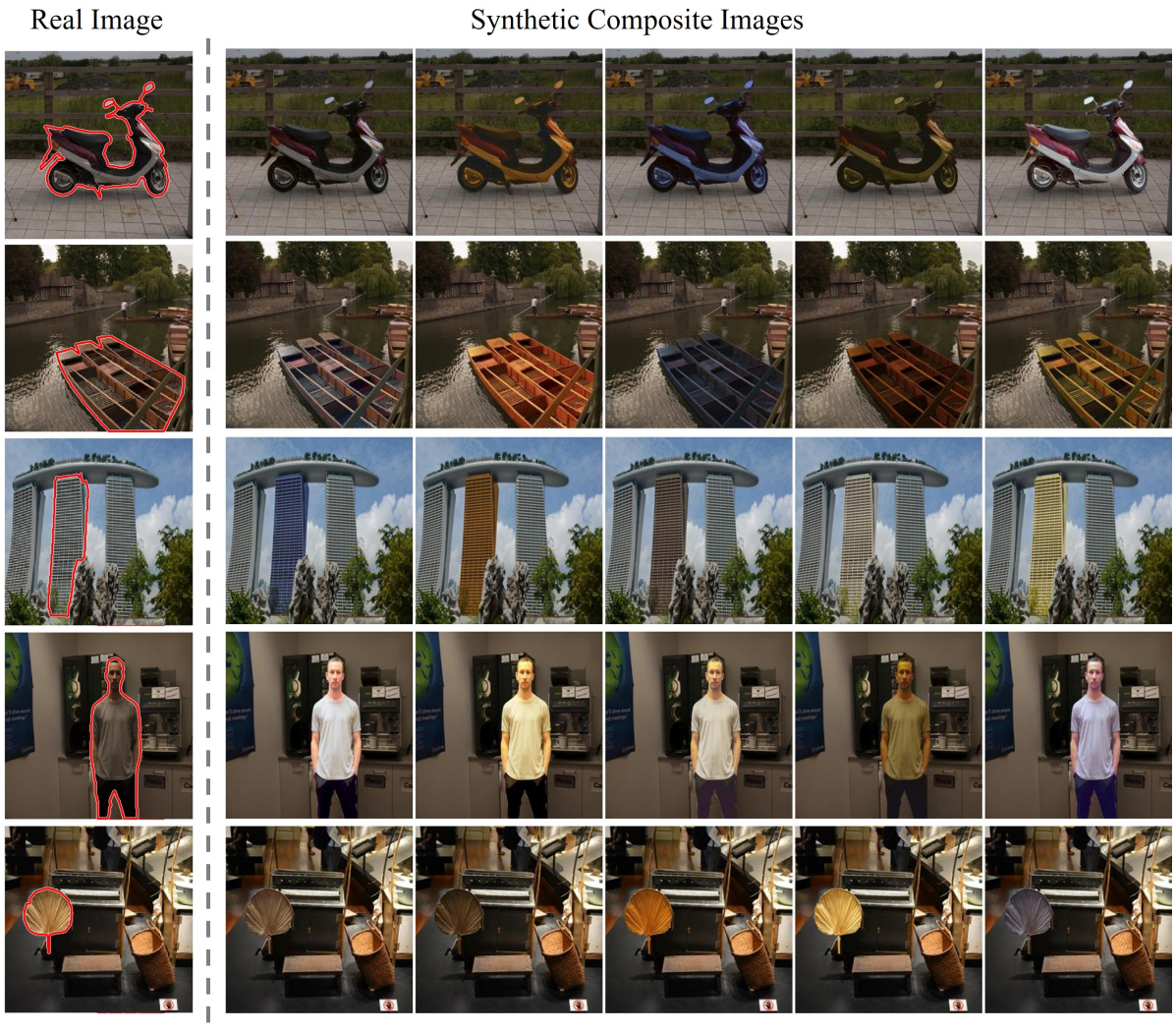

Figure 2. We show the real image (foreground outlined in red) in the leftmost column and five example synthetic composite images in the right columns.

Although the number of real images is limited by the existing datasets [1, 4], we argue that collecting images with color checker is more scalable than the way of constructing Hday2night [3], that is, using a fixed camera to capture the same scene over a long time span. Moreover, the main focus of this paper is exploring a novel way to construct harmonization dataset instead of building a large-scale dataset. This work has proved the feasibility of constructing harmonization dataset in this way, and ccHarmony could be easily extended by capturing more images with color checkers in the future.

1.2. Foreground Segmentation

For each real image, we manually segment one or two foregrounds. When selecting foregrounds, we ensure that the color checker can roughly represent the illumination information of the foreground, so that it is meaningful to apply the polynomial matching matrix calculated based on the color checker to the foreground. In total, we segment 426 foregrounds in 350 real images, in which the foregrounds cover a wide range of categories like human, tree, building, furniture, staple goods, and so on (see Figure 2).

1.3. Foreground Color Transfer

As described in Section 4 in the main paper, given an image $\boldsymbol{I}_a$, we first calculate its polynomial matching matrix $\boldsymbol{T}_a$ according to its color checker. Then, we apply $\boldsymbol{T}_a$ to the foreground $\boldsymbol{I}_a^f$ in $\boldsymbol{I}_a$ to convert it to $\boldsymbol{I}_o^f$, which is expected to be its counterpart in standard illumination condition. In Figure 1(c), we show several examples of real images $\boldsymbol{I}_a$ and their counterparts $\boldsymbol{I}_o$ in standard illumination condition.

Next, we randomly select 10 other real images as refer-

Figure 3. From left to right, we show the composite image (foreground outlined in red), the harmonized results of iSSAM [8], CDTNet [2], Harmonizer [6], DCCF [10], our GiftNet, and the ground-truth on iHarmony4 [3] dataset.

Figure 4. From left to right, we show the composite image, the foreground mask, the harmonized results of iSSAM [8], CDTNet [2], Harmonizer [6], DCCF [10], and our GiftNet on real composite images [2].

ence images for $\boldsymbol{I}_a$. For each reference image $\boldsymbol{I}_b$, we first calculate its inverse polynomial matching matrix $\boldsymbol{T}_b'$ according to its color checker. Then, we apply $\boldsymbol{T}_b'$ to $\boldsymbol{I}_o^f$ to convert it to $\boldsymbol{I}_b^f$, which is expected to be its counterpart in the illumination condition of $\boldsymbol{I}_b$.

Finally, we combine $\boldsymbol{I}_b^f$ and the background of $\boldsymbol{I}_a$ to form a synthetic composite image $\boldsymbol{I}_{a\to b}$. The above procedure has been introduced in the main paper, as illustrated in Figure 2 in the main paper. Since we select 10 reference images for each foreground, we can produce 10 synthetic composite images for each foreground. Based on 426 foregrounds, we can produce 4260 pairs of synthetic composite images and real images.

We split 350 images into 250 training images with 308 foregrounds and 10 test images with 118 foregrounds. Thus, the training set contains 3080 pairs of synthetic composite images and real images, while the test set contains 1180 pairs. We show several examples of real images and their corresponding synthetic composite images in Figure 2.

2. More Visual Comparison with Baselines

We provide more visualization results on iHarmony4 in Figure 3. Compared with the ground-truth real images, our method can usually produce harmonious and visually pleasing results, while the baseline methods may harmonize the composite images insufficiently or incorrectly.

3. Visual Results on Real Composite Images

The visual comparison between different methods on real composite images is shown in Figure 4. We can observe that our method can generally produce more harmonious and realistic images. For example, in row 1-2, the harmonized animals (penguin, panda) of our method are more similar to the same-category animals in the background. In row 3, the harmonized dog of our method is darker since it hides in the darkness. In row 4, the harmonized flower of our method is brighter with vivid color. In row 5-6, the harmonized cars of our method have more appealing and harmonious lustre. In row 7-8, the harmonized portraits of our method are more faithful to the background illumination, by referring to the people standing next to the foregrounds.

4. Visualization of Ablation Studies

Recall that we have conducted ablation studies for our method in Section 5.3 in the main paper. In Figure 5, we provide the visualization results of several ablated versions of our method, corresponding to row 1, 5, 8 in Table 3 in the main paper. From left to right, we observe that the harmonized images are getting closer to the ground-truth. The results of row 5 are better than those of row 1, demonstrating the effectiveness of our proposed GIFT module. The results of row 8 are better than those of row 5, which proves that relation distillation is useful.

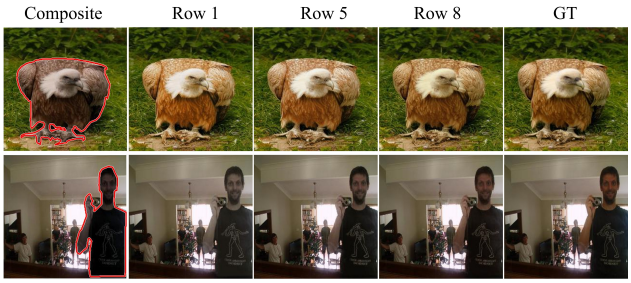

Figure 5. From left to right, we show the composite image, the harmonized result of our ablated versions (row 1, 5, 8 in Table 3 in the main paper), and the ground-truth.

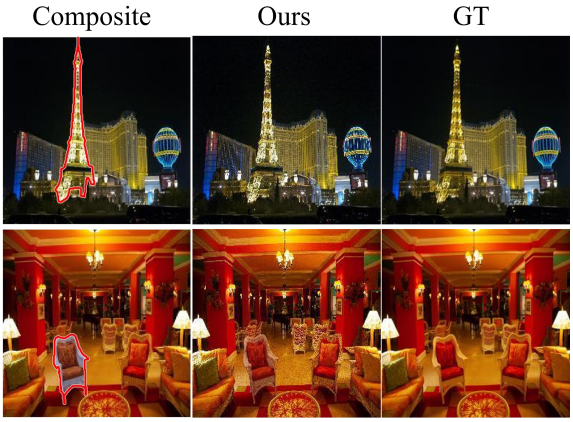

Figure 6. From left to right, we show the composite image, the harmonized result of our method, and the ground-truth.

5. Failure Cases

Although our method can usually generate satisfactory harmonized results, there also exist some cases where our method does not behave well. For example, as shown in Figure 6, for the artificial illumination sources, our method fails to produce satisfactory results, probably because most training images are with natural illumination sources.

Acknowledgement

The work was supported by the National Natural Science Foundation of China (Grant No. 62076162), the Shanghai Municipal Science and Technology Major/Key Project, China (Grant No. 2021SHZDZX0102, Grant No. 20511100300).

References

[1] Dongliang Cheng, Dilip K Prasad, and Michael S Brown. Illuminant estimation for color constancy: why spatial-domain

methods work and the role of the color distribution. *JOSA A*, 31(5):1049–1058, 2014. 1, 2

[2] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. *CVPR*, 2022. 3, 4

[3] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, 2020. 1, 2, 3

[4] Peter Vincent Gehler, Carsten Rother, Andrew Blake, Tom Minka, and Toby Sharp. Bayesian color constancy revisited. In *CVPR*, 2008. 1, 2

[5] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. In *ICCV*, 2021. 1

[6] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *ECCV*, 2022. 3, 4

[7] Xuqian Ren and Yifan Liu. Semantic-guided multi-mask image harmonization. In *ECCV*, 2022. 1

[8] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *WACV*, 2021. 3, 4

[9] Yazhou Xing, Yu Li, Xintao Wang, Ye Zhu, and Qifeng Chen. Composite photograph harmonization with complete background cues. In *ACM MM*, 2022. 1

[10] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *ECCV*, 2022. 3, 4