6. Supplementary material

In this supplementary, we first describe how our *Artist-Retouched* dataset was constructed, and show representative visual examples from the dataset (introduced in § 3.2). Then, we show the visual comparisons on iHarmony dataset [4] and report the detailed quantitative comparison results on four subsets (HCOCO, HAdobe5k, HFlickr, Hday2night). Besides, we provide more high-resolution visual results of different methods on the *Artist-Retouched* dataset and RealHM benchmark (supplementary to § 4.1). We then go in-depth into our real composite dataset with captured references and present more qualitative visual comparisons (supplementary to § 4.2). Furthermore, we show more visual comparisons of real composite images we used as part of our user studies (§ 4.2). Finally, we show more intermediate results and parametric outputs of our method for real-composite image harmonization (supplementary to § 4.3).

6.1. *Artist-Retouched* dataset

In this work, we propose to use a new *Artist-Retouched* dataset for our dual-stream training experiments. Unlike previous work, *Artist-Retouched* contains image pairs retouched by artists rather than mostly relying on random color augmentations. Artists were allowed to use global luminosity or color adjustments operations, but also local editing tools like brushes, e.g., to alter the shading. All the image editing was done using Adobe Lightroom, a software dedicated to photo adjustment. Figure S1 shows representative before-after image pairs in the *Artist-Retouched* dataset. *Artist-Retouched* consists of $n = 46173$ before/after retouching image pairs $\{\mathbf{I}_i, \mathbf{O}_i\}_{i=1,\dots,n}$, with the foreground mask $\mathbf{M}_i$ for each pair. As visualized in the figure, the retouching procedure consists of global luminosity/color adjustments (e.g., exposure, contrast, Highlights, Temp, Tint, Hue) and local editing tools(e.g., adding shading, creating soft transitions by gradient mask). From each triplet $\{\mathbf{I}_i, \mathbf{O}_i, \mathbf{M}_i\}$, we can generate two synthetic composites inputs for training: one with only the foreground retouched $\mathbf{M}_i \cdot \mathbf{O}_i + (1 - \mathbf{M}_i) \cdot \mathbf{I}_i$, and the other one with only the background being retouched $\mathbf{M}_i \cdot \mathbf{I}_i + (1 - \mathbf{M}_i) \cdot \mathbf{O}_i$. We use the unedited image $\mathbf{I}_i$ and the retouched image $\mathbf{O}_i$ as ground truth targets of these composite inputs, respectively.

6.2. More results on iHarmony benchmark

As discussed in §4.1, we evaluate our method on the iHarmony benchmark [4] and present the quantitative results on the entire dataset. In this section, we report the quantitative results on four subsets of iHarmony — HCOCO, HAdobe5k, HFlickr, Hday2night. We compare our method with DovNet [4], IHT [9], Harmonizer [14]. Our method outperforms or matches state-of-the-art approaches in all four subsets of iHarmony benchmark. Table S1 summarizes the quantitative results. Besides, Figure S2 shows a gallery of selective visual comparisons between different approaches at 512×512 resolution. For better visualization, we resize the images to their original aspect ratios.

6.3. More visual results on *Artist-retouched* dataset

As introduced in § 3.2, we evaluate different methods on a testing split of our *Artist-Retouched* dataset with realistic retouches from human experts. In addition to the results shown in Figure 5, Figure S3 presents more visual comparisons on *Artist-Retouched* testing dataset at 1024 resolution. We observe that our results agree better with the ground truth images in terms of the visual quality compared to other methods (DovNet [4],IHT [9],Harmonizer [14]). Besides, we also show one failure example (boat), where all methods (including ours) fail to retrieve the correct color of the ground truth image, though some of them look harmonious by themselves without seeing the reference. We hypothesize that, in this case, the skylight illumination in the ground-truth image is difficult to infer from the background.

6.4. More visual results on *RealHM* benchmark

Different from synthetic dataset [4], RealHM [13] benchmark contains 216 real-world high-resolution composites with expert annotated harmonization results as ground truth. In this section, we present more visual harmonization comparisons in Figure S4 at 1024 resolution. From the results, we observe that even though there exist strong foreground/background color mismatches in the real composite images, our method produces more harmonious results compared to other approaches.

6.5. Real composites with captured reference

As briefly introduced in § 4.2, for qualitative evaluation, we created a dataset of 40 high-resolution real-composite images with captured references. As illustrated in Figure S5, we first capture a fixed set of foreground objects against multiple backgrounds (scenes), as well as the corresponding "background-only" images. We then segment the foreground object of one photo and paste it onto the "background-only" photo of another with roughly the same location. The captured photo of the same object in the same background scene serves as a reference for qualitative evaluation. Figure S6 visualize selective examples of the harmonization results. We compare our method with state-of-the-art approaches. As shown in the figure, for the first example, our result shows better visual agreements with the captured reference. For the second and third examples, though our results don't exactly match the reference (none of the other methods does), our method still produces images with harmonious appearances. We will release this

dataset upon publication.

6.6. More results on real composite images

Figure S7 shows more visual comparisons on real composite images where we don't have the ground truth or captured reference. We use these images as part of our user studies (§4.2). We compare our method with DovNet [4], IHT [9], Harmonizer [14]. We will release these testing real composite images upon the publication of this work.

6.7. More intermediate results

Figure S8 presents more intermediate results and parametric outputs on RealHM [13] real-composite benchmark. As shown in the results, our predicted RGB curves harmonize the global color/tone, while our learned shading map incorporates local shading to the final outputs. By comparing with the human-annotated harmonization results (right), we observe that our local shading maps align well with the local operation done by the human experts. For instance, for the top three rows, both our results and the human-annotated ground truth selectively darkened the bottom part of the foreground objects. For the fourth row example, our result highlights the region with incoming light while darkening other foreground parts, which agrees with the operations done by human experts.

6.8. Demo video

To further better demonstrate the effectiveness of our method in real-world applications, we prepared and recorded a demo video (see attachments of the supplementary material). We can interactively run our demo on a single CPU without access to extensive computing resources.

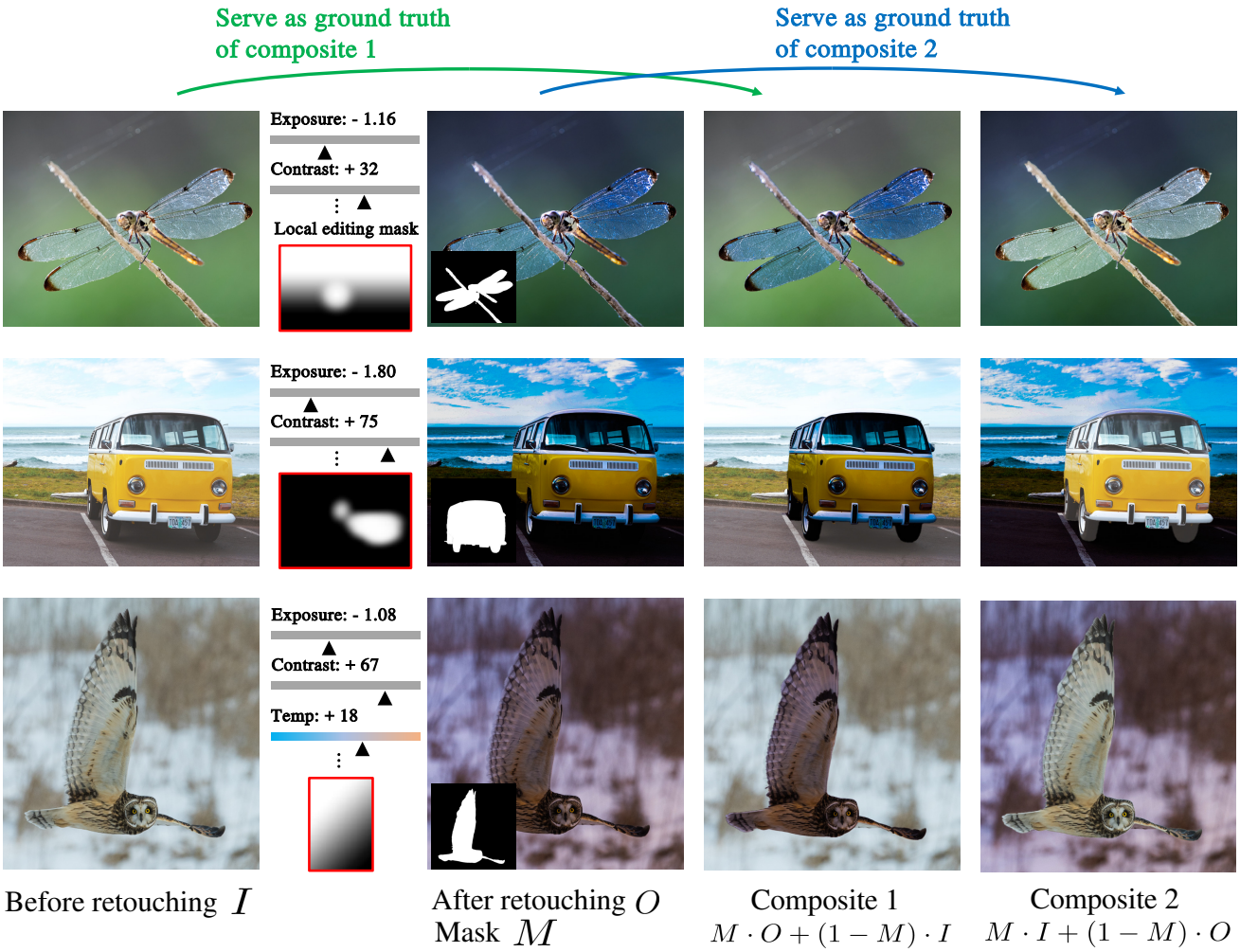

Figure S1. **Construction of *Artist-Retouched* dataset.** *Artist-Retouched* dataset contains before/after artist-retouching image pairs $\{\mathbf{I}_i, \mathbf{O}_i\}_{i=1,\ldots,n}$ with the foreground mask $\mathbf{M}_i$ for each pair. Artist retouching procedures include both global luminosity/color adjustments as well as local editing. Local editing masks (images with red borders) in the figure indicate the selective regions where artists perform local editing (e.g., shading). Two composite images (Composite 1 and 2) are created and used for training from each pair of images. Unedited image I and retouched image O serve as the ground truth for composite 1 and 2, respectively.

Method	HCOCO		Adobe5k		HFlickr		Hday2night		Entire dataset	
	PSNR↑	SSIM ↑	PSNR↑	SSIM ↑	PSNR↑	SSIM↑	PSNR↑	SSIM ↑	PSNR↑	SSIM ↑
Composite	33.92	0.9862	28.51	0.9563	28.44	0.9638	34.32	0.9741	31.74	0.9748
DovNet [4]	35.76	0.9875	35.05	0.9733	30.68	0.9711	34.83	0.9707	34.97	0.9812
IHT [9]	38.38	0.9924	37.02	0.9819	32.84	**0.9810**	36.79	0.9763	37.33	0.9877
Harmonizer [14]	**38.77**	**0.9936**	**38.97**	**0.9888**	**33.71**	**0.9833**	**37.96**	**0.9813**	**38.25**	**0.9909**
Ours	**39.07**	**0.9940**	**38.53**	**0.9835**	**33.60**	0.9793	**38.15**	**0.9817**	**38.30**	**0.9891**

Table S1. **Quantitative comparisons on subsets of iHarmony benchmark [4]** at 256×256 resolution. We compare our method with DovNet [4], IHT [9], Harmonizer [14]. PSNR and SSIM are used as metrics. **Red**, and **Blue** correspond to the first and second best results. ↑ means higher the better, and ↓ means lower the better.

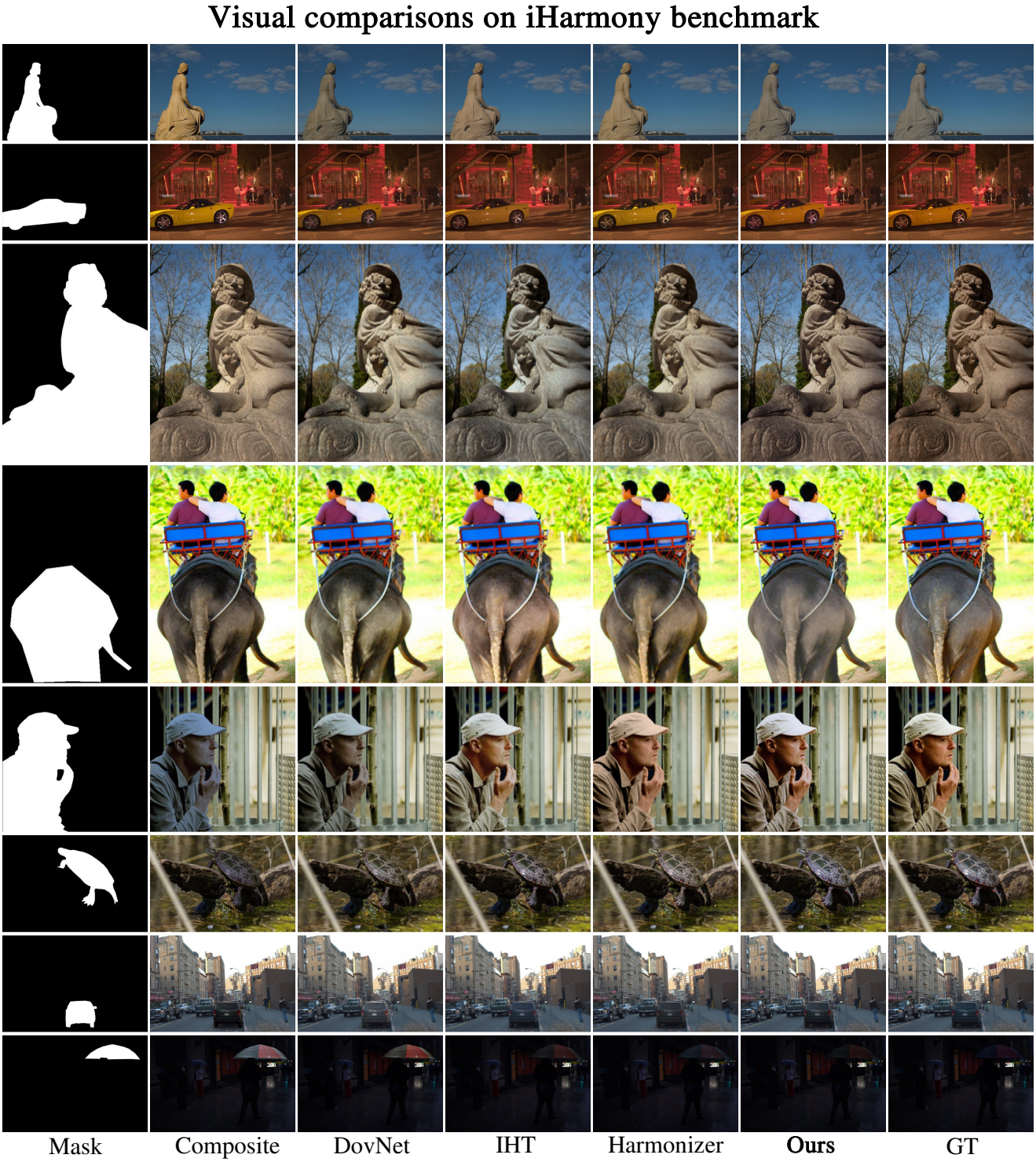

Figure S2. **Representative visual comparisons between state-of-the-art harmonization methods on iHarmony benchmark.** We compare our method with composite image, DovNet [4], IHT [9], Harmonizer [14], and ground truth. Foreground masks are displayed in the first column. For better visualization, we resize the images to the original aspect ratio. Our method shows better visual alignment with the ground truth images than other state-of-the-art methods.

Figure S3. **More visual comparisons between state-of-the-art harmonization methods on *Artist-Retouched* dataset.** We compare our method with composite image, DovNet [4], IHT [9], Harmonizer [14], and ground truth. Red boxes indicate the foreground mask of the composite images. Our method shows better visual alignment with the ground truth images compared to other state-of-the-art methods. We also present one failure example, where all methods fail to recover the ground truth appearance, though some of them look harmonious without referring to the ground truth.

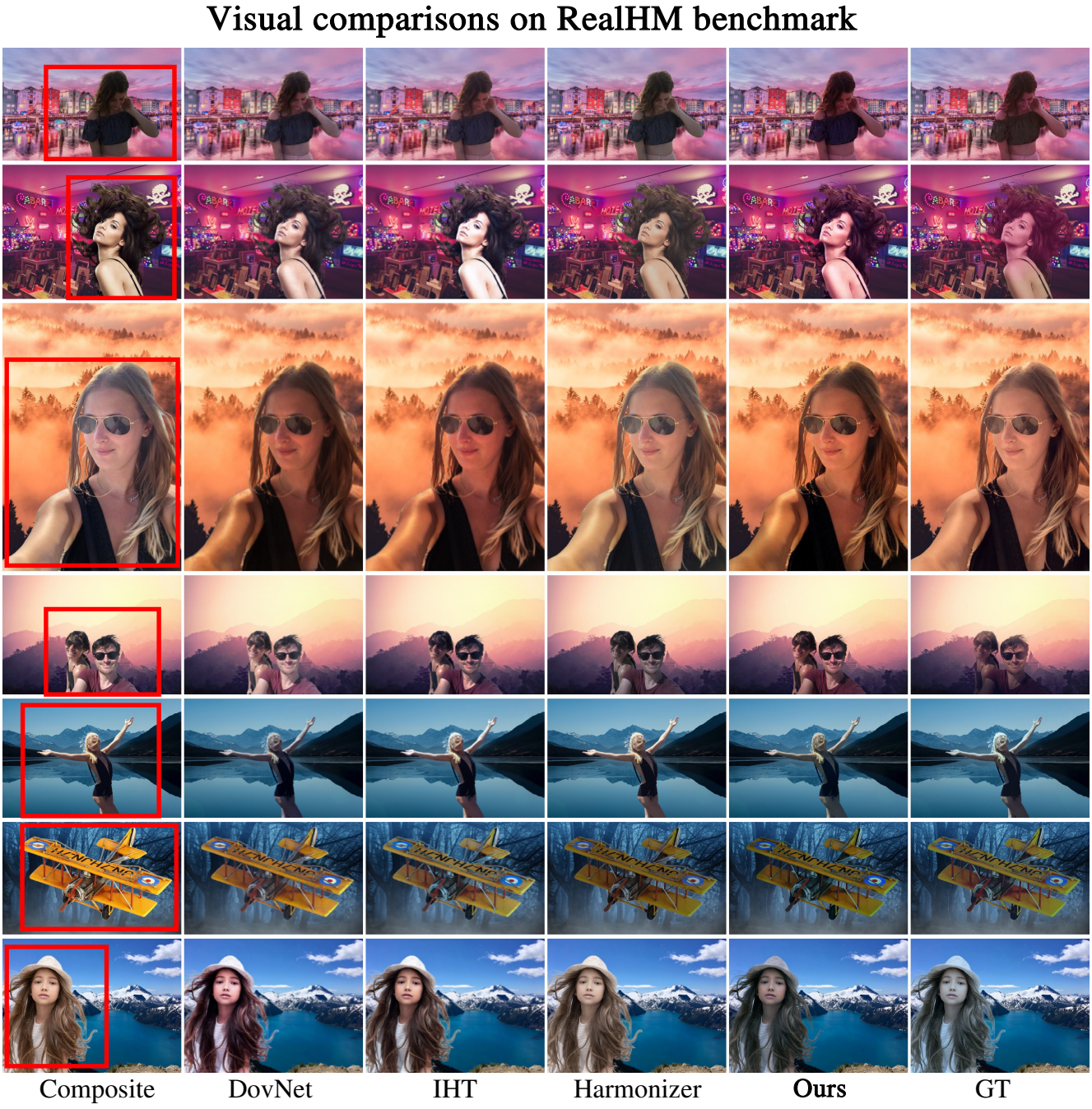

Figure S4. **More visual comparisons between state-of-the-art harmonization methods on RealHM benchmark.** We compare our method with composite image, DovNet [4], IHT [9], Harmonizer [14], and ground truth. Our method shows better color consistency with the ground truth images (row 1, 2, 4, 6, and 7) and deliver more harmonious results.

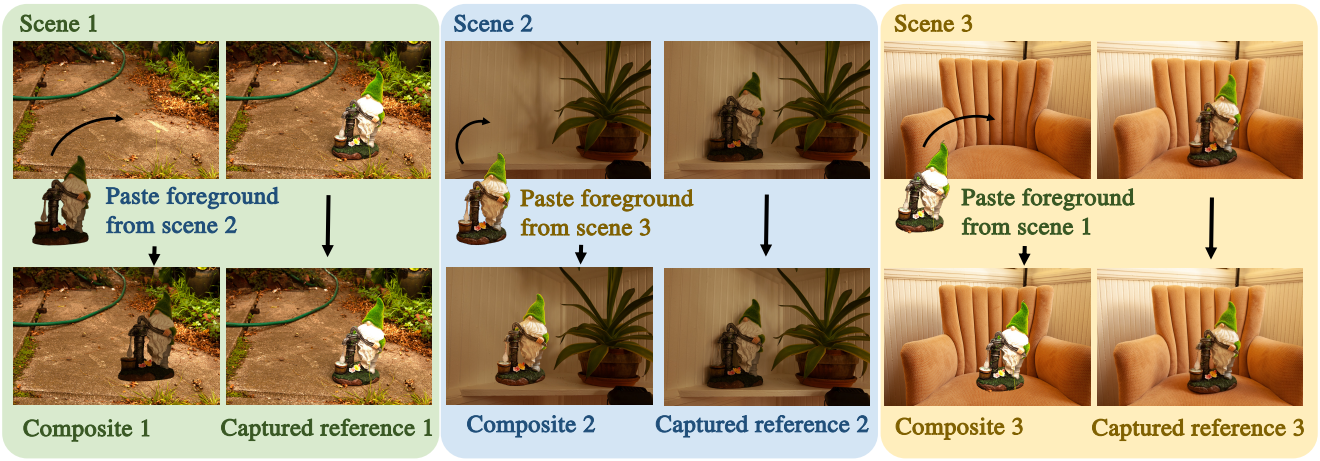

Figure S5. **Construction of composite images with captured reference.** First, we capture the same foreground object against multiple backgrounds (3 backgrounds in the figure), as well as the corresponding "background-only" photos. We then segment the foreground object from one photo and paste it onto the "background-only" image of another to generate the composite images. The captured photo of the same object in the same background scene serves as qualitative references (Here, captured references 1, 2, and 3).

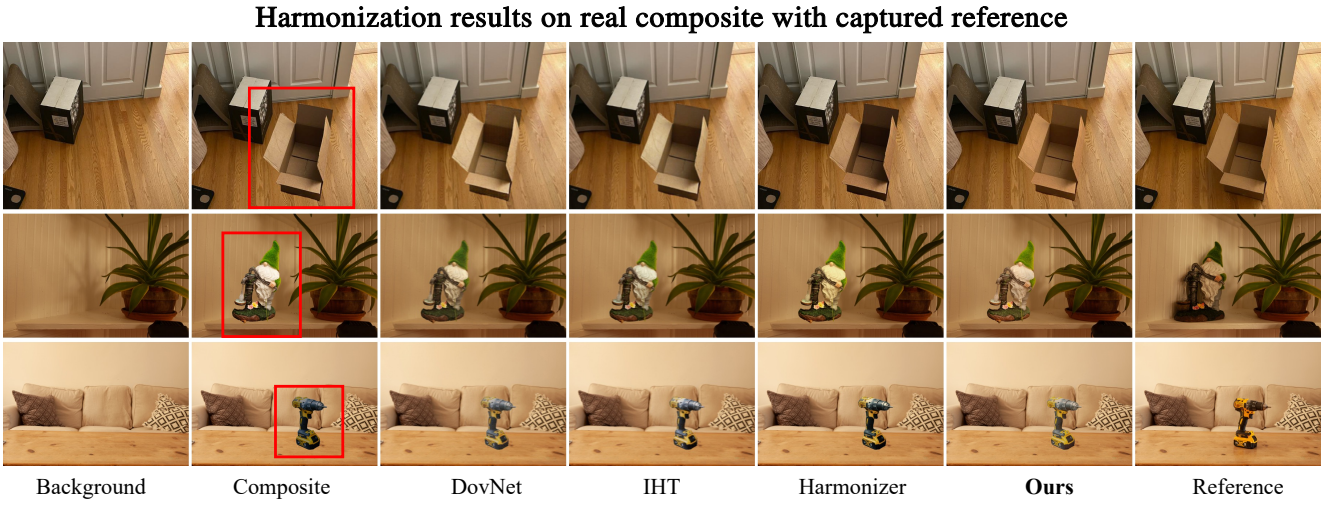

Figure S6. **Real composite harmonization results with captured reference.** The composite is obtained by pasting the foreground object from a different photo (not shown) onto the background (left). The reference (right) is obtained by physically placing the foreground object in the background scene and taking a photo. We compare our method with composite image, DovNet [4], IHT [9], Harmonizer [14], and the captured reference.

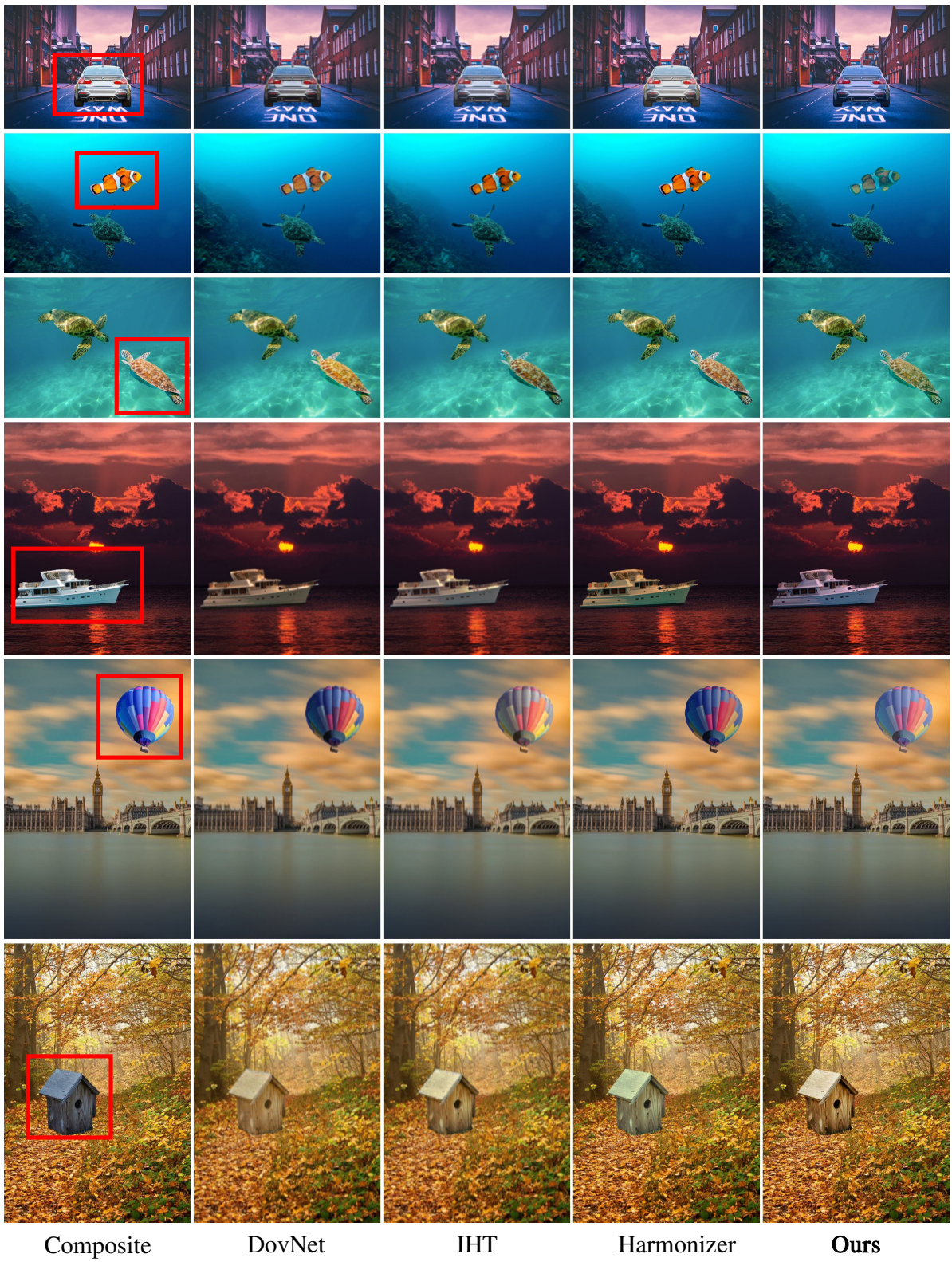

Figure S7. **More visual comparisons on real composite images.** We compare our method with composite image, DovNet [4], IHT [9], and Harmonizer [14].

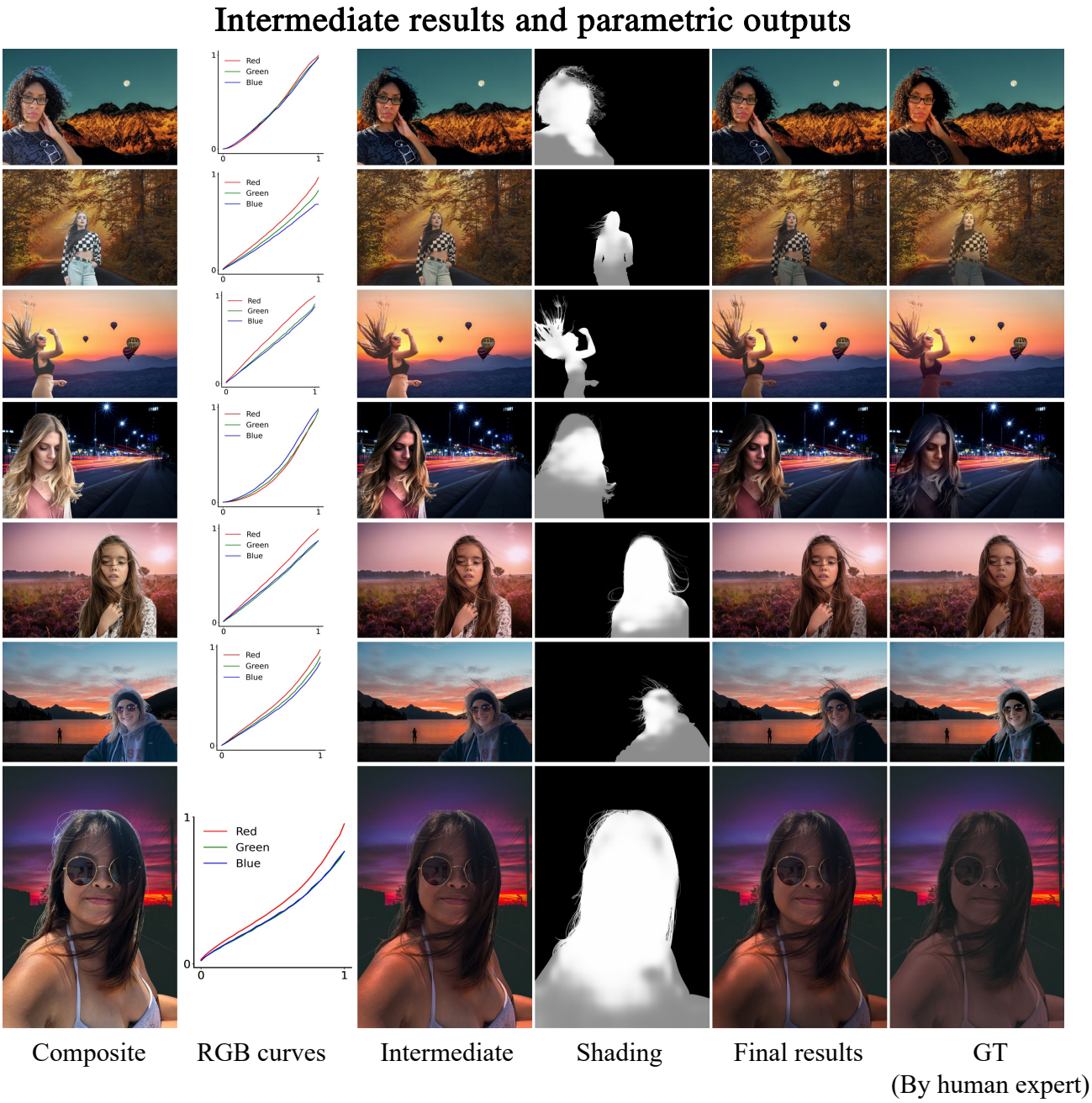

Figure S8. **Intermediate results and parametric outputs on RealHM benchmark.** RGB curves harmonize the global color/tone (third column), while our shading map corrects the local shading in the final harmonization outputs (fifth column). Our local shading maps agree well with the local shading operations done by human experts/artists (right column).