# Supplementary Material for Semantic-guided Multi-Mask Image Harmonization

Xuqian Ren[1[0000−0002−3811−0235]], Yifan Liu[2[0000−0002−2746−8186]]

Beijing Institute of Technology
University of Adelaide

In this supplementary material, we will first show more illustration cases about the explainable property of our proposed framework. Second, we show more qualitative results of baseline and our proposed framework. Third, more visualizations in our user study will be provided. Next, some ablation studies is conducted. Finally, we will compare our framework with other state-of-art methods on a public benchmark with single masks.

## 1 Explainable property of our framework

Here we show more results to demonstrate the explainable property of our framework by visualizing operator masks. Figure 2 are the results from HScene, and Figure 3 shows results from HLIP. We can see that our framework can sense the disturbance in various directions and give a reasonable explanation for its adjustment direction. Thus, our method can make the harmonization process interpretability as well as ensure the performance is similar to previous baselines. We can also see that our framework can harmonize two completely different adjustment directions in an image (like results show in the 1, 7, 8, 9, $10th$ line of Figure 2 and the 1, 2, 5, 7, 8, 9, $10th$ line of Figure 3), which will not make the framework adjust the foreground in only one direction, so our framework is foreground-aware network.

## 2 Qualitative results between different output formats

In Figure 4, we show more qualitative visualization results between the RGB output format and our operator masks output format. Our operator mask-based framework is more suitable for structural adjustment, even there are different harmonization directions in one image. Also, training in a GAN framework can make the model pay more attention to the details(as can be seen in the $9th$ line in Figure 4).

## 3 More User study results

Here we show more pictures used in the User Study in Figure 1. The first three lines show results that only illumination needs to be changed. The 4, $5th$ shows the results when foregrounds have already been processed by Instagram filters.

The last few lines provide the results when foregrounds from different pictures, but the composited images already look real. So our model tends to keep them almost unchanged.

## 4 Ablation study on HScene

In this section, we show the effect of each component in our framework in Table 1. It can be seen that both the perceptual loss and operator masks add improvement to the final harmonious image. In order to better study the localization of our framework, we limit the output and changes to $OM_{add}$ and binarize the operator masks. The IOU between our OMs and ground truth masks is 52.9%.

**Table 1.** Effects of different components in our framework.

| $OM_{add}$ | $OM_{mul}$ | $\mathcal{L}_{Lpips}$ | $D$ | MSE↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|---|---|---|---|
| | | ✓ | | 141.75 | 28.73 | 0.95 | 0.026 |
| ✓ | | ✓ | | 92.45 | 30.79 | 0.96 | 0.052 |
| ✓ | ✓ | ✓ | | 99.84 | 30.59 | 0.96 | 0.020 |
| ✓ | ✓ | | | 105.27 | 30.29 | 0.96 | 0.021 |
| ✓ | ✓ | ✓ | ✓ | 94.00 | 30.88 | 0.96 | 0.020 |

## 5 Two-Stage v.s One-Stage

Our framework can harmonize regions without input masks, integrating localization and harmonization functions. We compare our one-stage pipeline with a two-stage pipeline. Previous methods need masks as input. When masks are not provided, the localization method needs to be first used to get the inharmonious region. We first use DIRL[4] to locate the inharmonious regions and then use DoveNet[2] to harmonize them progressively. The results can be seen in Table 2. Our single-stage model greatly saves the processing time and have higher precision.

**Table 2.** Comparison between the Two-Stage pipeline and the One-Stage pipeline. The time is the whole seconds to process the test images. ‡ means we fine-tune the whole model. We fine-tune DIRL 60 epochs and DoveNet 200 epochs.

| Methods | HScene | | | | | HLIP | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MSE↓ | PSNR↑ | SSIM↑ | LPIPS↓ | Time(s)↓ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | Time(s)↓ |
| DIRL‡[4]+DoveNet‡[2] | 158.94 | 29.06 | 0.95 | 0.066 | 287.6 | 80.34 | 31.45 | 0.96 | 0.047 | 1430.0 |
| Sg-MMH(OMGAN) | 94.00 | 30.88 | 0.96 | 0.020 | 106.6 | 42.62 | 33.92 | 0.97 | 0.018 | 599.0 |

## 6 Results on iHarmony4

Our framework mainly focuses on structural perturbations and the multi-mask harmonization task. However, to further verify the effectiveness of our framework, we tested it on the existing dataset. We choose iHarmony4[2] to further

explore the ability of our framework when it is applied to a normal single-mask harmonization task. We adapted an Harmonization Transformer [3] (HT) to implement our framework and replaced the output as operators masks.The results are shown in in Table 3.

In order to explore the effects of more kinds of operation masks, we also implement experiments on HLS Color space, for it has more decoupled channels. The operations on the H channel can manipulate the color independently, and the operator masks on L and S channels can change the illumination and saturation individually. So it is a better space to manipulate pictures more intuitively. We visualize the explainable and editable properties in HLS color space with control of the output with the simplest $OM_{add}$ operator mask. In Figure 5, we visualize the $OM_{add}$ in H, L, and S channels, and in Figure 6, we change the value of the $OM_{add}$ in each channel separately to see the editable attribute of the operator masks. We also binarize $OM_{add}$ with a threshold of 1e-4 and further calculate the mask IOU of $OM_{add}$ with ground truth masks (51.1% for HCOCO, 85.3% for HAdobe5k, 81.5% for HFlickr, 77.9% for Hday2night.)

**Table 3.** Quantitative results on four sub-datasets of iHarmony4. **Bold** means the best results with our output format, **Blod** means the next best result. ‡ means we re-train the whole model.

| Sub-dataset | HCOCO | | HAdobe5k | | HFlickr | | Hday2night | | All | |
|---|---|---|---|---|---|---|---|---|---|---|
| Evaluation Metric | MSE↓ | PSNR↑ | MSE↓ | PSNR↑ | MSE↓ | PSNR↑ | MSE↓ | PSNR↑ | MSE↓ | PSNR↑ |
| Composite | 69.37 | 33.94 | 345.54 | 28.16 | 264.35 | 28.32 | 109.65 | 34.01 | 172.47 | 31.63 |
| DoveNet(RGB)[2] | 36.72 | 35.93 | 52.32 | 34.34 | 133.14 | 30.21 | 54.05 | 35.18 | 52.36 | 34.75 |
| HT(RGB‡)[3] | 14.64 | 39.04 | 22.03 | 38.85 | 54.34 | 34.12 | 52.26 | 36.75 | 21.91 | 38.39 |
| HT(HLS, $OM_{add}$) | 21.60 | 37.45 | 29.48 | 36.89 | 78.29 | 32.59 | 63.81 | 36.36 | 31.00 | 36.73 |
| HT(HLS, $OM_{add} + OM_{mul}$) | 21.47 | 37.56 | 29.17 | 37.15 | 73.96 | 32.83 | 55.11 | 36.84 | 30.19 | 36.90 |
| HT(LAB, $OM_{add} + OM_{mul}$) | **13.95** | **39.14** | **22.04** | **38.64** | **59.03** | **34.00** | **43.57** | **36.86** | **21.89** | **38.38** |

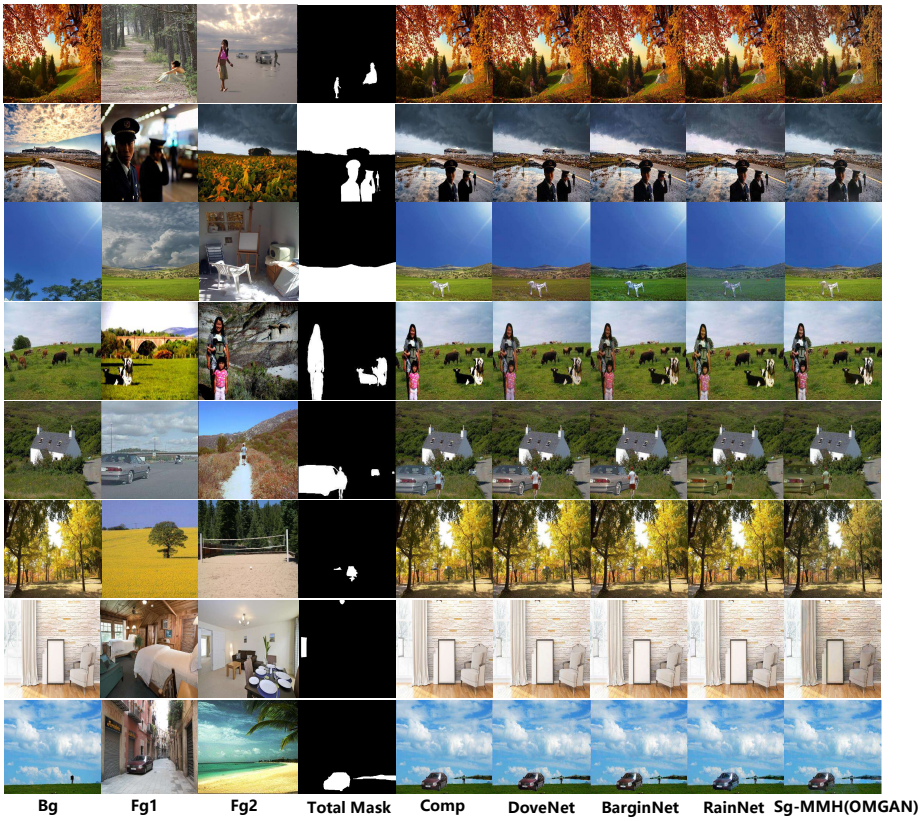


**Fig. 1.** Here we show more visualizations used in the User Study to compare the results of our method with DoveNet[2], BarginNet[1], and RainNet[5].

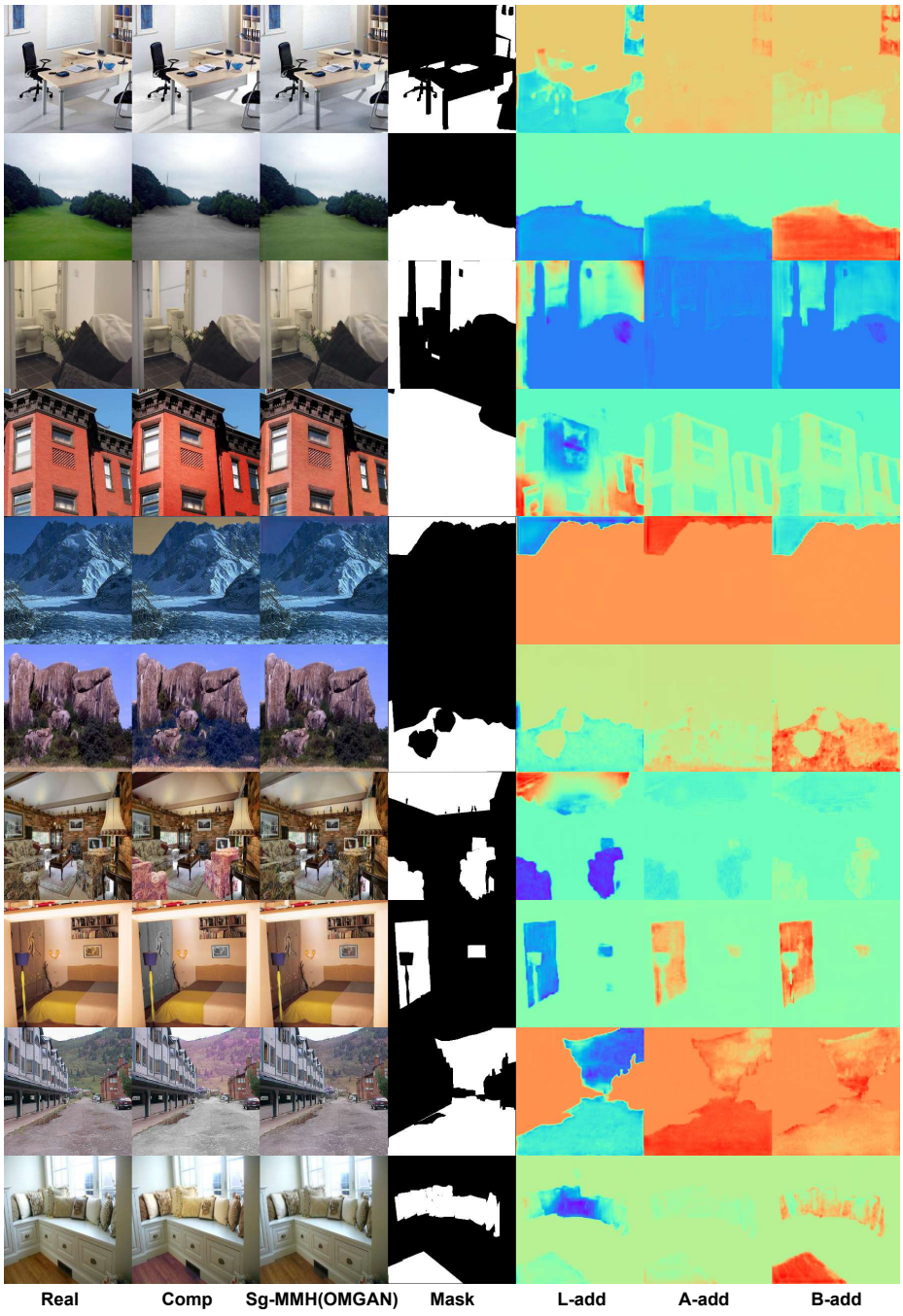


**Fig. 2.** Here we visualize more examples from HScene. Our Operator Masks can make positive or negative responses according to different changes in the foregrounds.

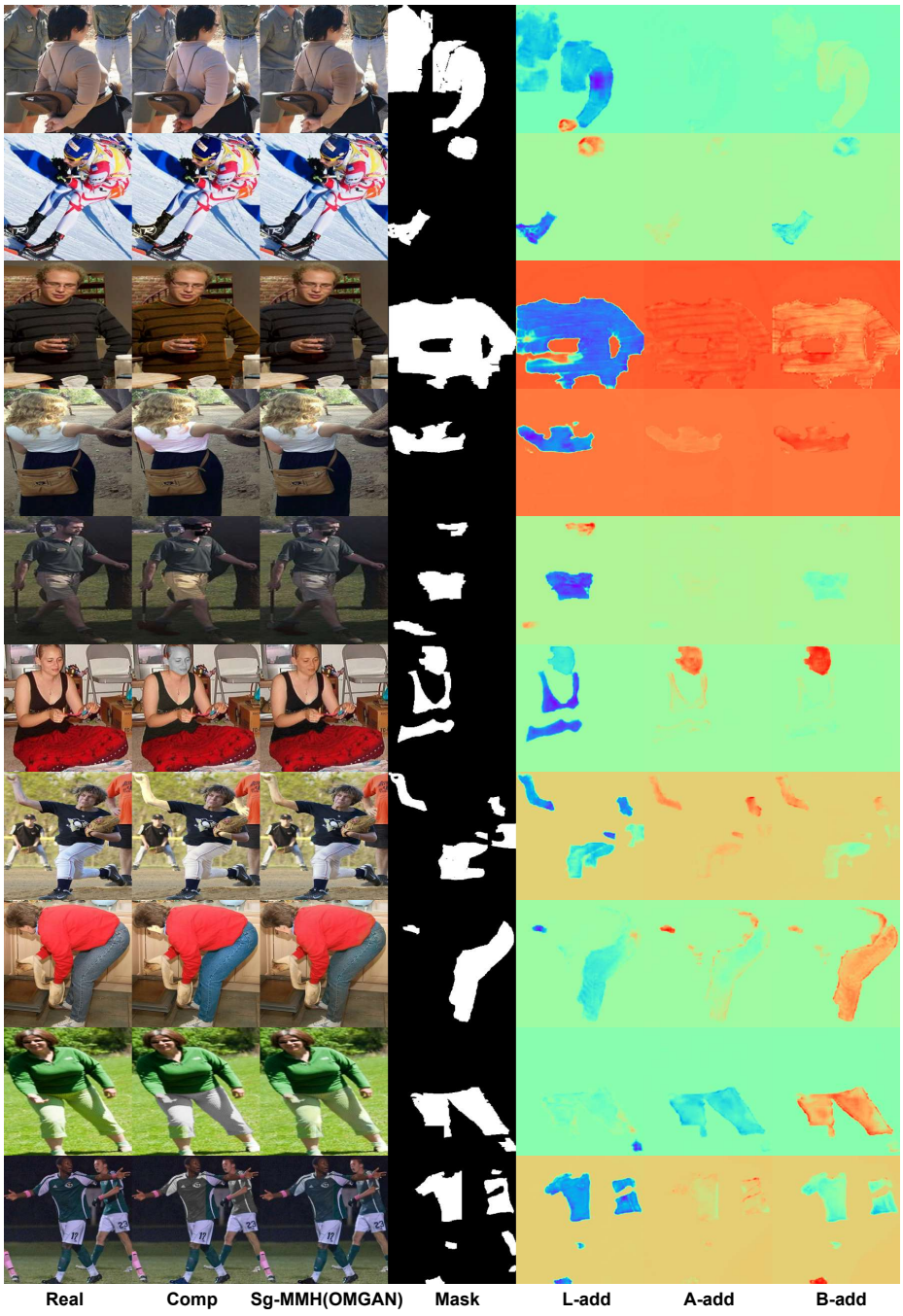


**Fig. 3.** Here we visualize more examples from HLIP. Our Operator Masks can make positive or negative responses according to different changes in the foregrounds.

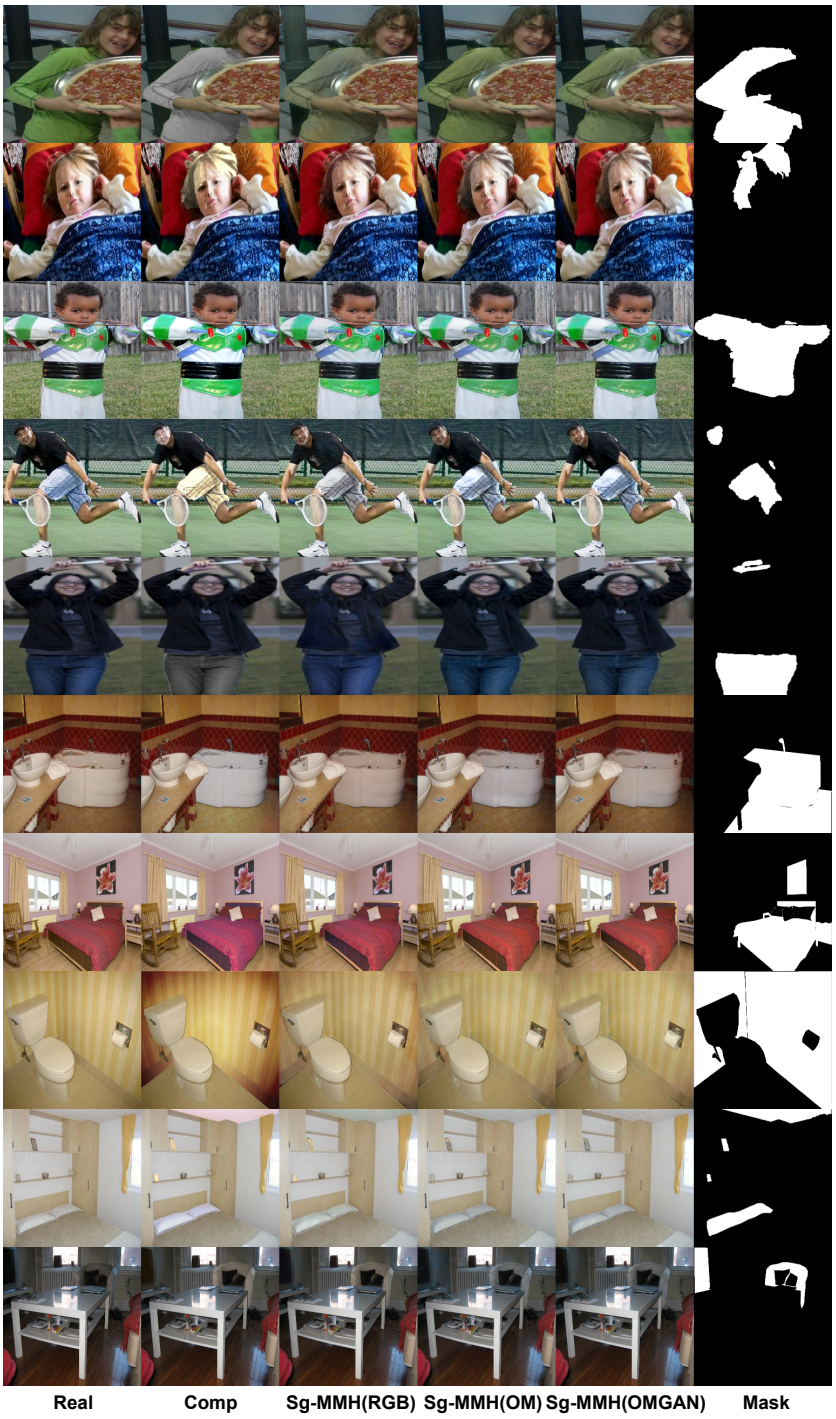


**Fig. 4.** Here we show more comparison results between RGB output format and our Operator Mask output w/wo discriminator.

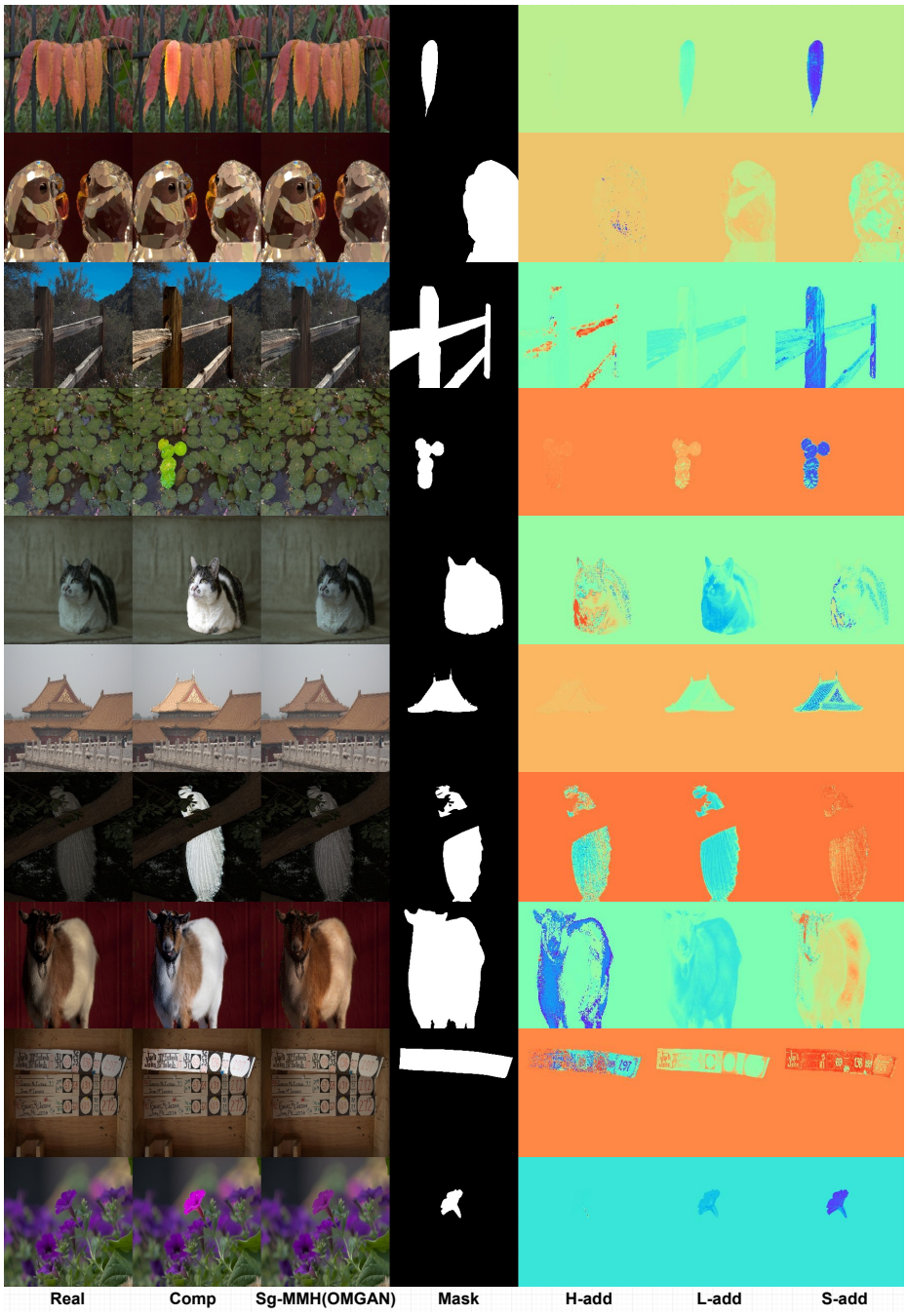


**Fig. 5.** Here we visualize some examples from iHarmony4. Our Operator Masks can make positive or negative responses according to different changes in the foregrounds.


**Fig. 6.** Here we visualize the editable property with samples from iHarmony4. We change the $OM_{add}$ in H, L, and S channels separately to gain various outputs.

## References

1. Cong, W., Niu, L., Zhang, J., Liang, J., Zhang, L.: Bargainnet: Background-guided domain translation for image harmonization. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
2. Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., Zhang, L.: Dovenet: Deep image harmonization via domain verification. In: CVPR. pp. 8394–8403 (2020)
3. Guo, Z., Guo, D., Zheng, H., Gu, Z., Zheng, B., Dong, J.: Image harmonization with transformer. In: ICCV. pp. 14870–14879 (2021)
4. Liang, J., Niu, L., Zhang, L.: Inharmonious region localization. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
5. Ling, J., Xue, H., Song, L., Xie, R., Gu, X.: Region-aware adaptive instance normalization for image harmonization. In: CVPR. pp. 9361–9370 (2021)