LEMaRT: Label-Efficient Masked Region Transform for Image Harmonization

Supplementary Material

1. Qualitative Examples

We present additional qualitative examples of the output of our method (LEMaRT [SwinIH]) and three SOTA methods (RainNet [10], iS^{2}AM [12], DHT+ [5]) on the iHarmony4 dataset in Figure 3 and Figure 4. Similar to what we observe in Figure 5 of the main paper, our method is better at color correction. The images generated by our method have more natural colors and are closer to the ground truth images. We also provide qualitative examples of the output of our method (LEMaRT [SwinIH]), iS^{2}AM [12] and DHT+ [5] on the RealHM dataset in Figure 5. We see from the first five examples that our method better harmonizes composite images. We show a controversial example in the last row. Different people may have different opinions regarding which harmonized image looks more natural.

2. Details of Data Generation Pipeline

The data generation pipeline of LEMaRT uses ten different transformations to generate the data for pre-training. The transformations include adjustments to brightness, contrast, hue, saturation and sharpness as well as blurring, deblurring, auto contrast, equalization and posterization. The first five transformations adjust the brightness, contrast, hue, saturation and sharpness of an image by a factor of c, respectively, $c \in [0.2, 1.8]$ for brightness adjustment; $c \in [0.3, 1.7]$ for contrast adjustment; $c \in [0.7, 1.3]$ for hue adjustment; $c \in [0.5, 1.5]$ for saturation adjustment; $c \in [0.0, 2.0]$ for sharpness adjustment. We sample c uniformly. For blurring, Gaussian blur with kernel size (k_1, k_2) ($k_1 \in [3, 9], k_2 \in [5, 11]$) is applied. For deblurring, we apply Gaussian blur to an image. The *blurred image* is treated as the original image and the unblurred image is treated as the transformed image. Auto contrast maximizes the contrast of an image by remapping its pixel values so that the lowest value becomes 0 and the highest value becomes 255. Equalization adjusts the histogram of an image so that the histogram of the output image has a uniform grayscale distribution. Posterization reduces the number of bits for each color channel of an image to n bits. n is uniformly sampled from $\{1, 2, 3, 4, 5, 6\}$.

In Figure 2, we present additional examples that show data generated for pre-training, *i.e.*, the transformed images, the masks and the composite images, and the output of a pre-trained LEMaRT model given the generated data (please refer to Figure 2 of the main paper for an illustration of the data generation pipeline of LEMaRT). The seven different transformations used to perturb the original images are attached to the transformed images. We can see that, after being pre-trained, our LEMaRT model can harmonize the composite images that are generated using a variety of transformations, *e.g.*, brightness adjustment, posterization, blurring. The results show that the our LEMaRT model learns to handle different factors that cause appearance mismatch between the foreground and the surrounding background.

dataset	metric	transformation diversity: standard	transformation diversity: less
all	PSNR↑	39.0	38.2
	MSE↓	20.9	26.0
	fPSNR↑	26.6	25.7
	fMSE↓	250.0	301.2

Table 1. Comparison of the performance of our method pre-trained using different transformations.

To understand the impact of the diversity of transformations to the performance of our model (LEMaRT [SwinIH]), in Table 1, we compare the performance of our model pre-trained with the set of transformations introduced above (denoted as standard) and a set of transformations with less diversity (denoted as less). Specifically, we halve the range of the factor c that controls the diversity of the five transformations which adjust the brightness, contrast, hue, saturation and sharpness of an image, *i.e.*, $c \in [0.6, 1.4]$ for brightness adjustment; $c \in [0.65, 1.35]$ for contrast adjustment; $c \in [0.85, 1.15]$ for hue adjustment; $c \in [0.75, 1.25]$ for saturation adjustment; $c \in [0.5, 1.0]$ for sharpness adjustment. We sample c uniformly. For blurring, Gaussian blur with kernel size (k_1, k_2) ($k_1 \in [2, 5], k_2 \in [2, 5]$) is applied. We do not use equalization. We see from Table 1 that pre-training our model using transformations with less diversity results in 0.8 dB drop in PSNR and 5.1 increase in MSE. This indicates that the diversity of the transformations has direct influence on the performance of our model. However, empirically, we find that increasing the diversity of the transformations further does not lead to better performance. A possible reason is that samples created by those

transformations are so unnatural that they seldom appear in real world.

3. Comparison with SOTA Methods

We compare our method, LEMaRT [SwinIH], with SOTA methods on iHarmony4. The results are shown Table 3. Table 3 differs from the Table 1 of the main paper only in that it shows all four metrics, *i.e.*, PSNR, MSE, fPSNR, fMSE (due to space constraints, the Table 1 of the main paper does not show fPSNR and fMSE). Similar to what we observe from the Table 1 of the main paper, our method consistently outperforms other methods across the two additional metrics (fPSNR and fMSE) on all subsets of iHarmony4. Our method achieves a fPSNR of 27.2 dB, which is 1.3 dB higher than the previous best method. The fMSE of our method is 213.3, which is 35.6 lower (14.3% relative improvement) than the previous best method [7].

In Table 4, we further compare our method with four additional image harmonization methods [2, 8, 9, 11] for completeness. [8, 9] are published in ECCV'22. [11] is an arXiv paper and [2] is a CVPR'22 paper. These four methods *underperform* SCS-Co [7] with which we compare our method in Table 1 of the main paper. We only present PSNR and MSE as all four methods do not report fPSNR and three of them, *i.e.*, [2, 9, 11], do not report fMSE. We see that our method outperforms all four methods. Our method achieves a PSNR of 39.8 dB which is 1.6 dB higher than [11], *i.e.*, the best of the four methods. The MSE of our method is 7.2 lower (30.0% relative improvement) than [11]. Harmonizer also adopts a perturbation-reconstruction strategy for training data generation. While Harmonizer [8] applies transformations to perturb the manually labeled foreground, LEMaRT perturbs regions specified by automatically generated masks. LEMaRT generates training data automatically using the plentiful supply of unlabeled data. We see that the MSE of our method is 8.8 higher than that of Harmonizer. It is likely that this is caused by a few images on which our method performs poorly (largest MSE over 1200.0, more than 27 times of the average). As PSNR is in log space, the images on which our method performs poorly has less influence to PSNR than MSE.

4. Cause of Block-shaped Artifacts

In Figure 7 of the main paper, we show that only using shifted window (Swin) attention *occasionally* causes block-shaped artifacts. We explain the reason why the block-shaped artifacts appear. In Figure 1, we present an illustration of the Swin attention. Visual tokens within the same window (shown in the same color) can attend to each other, but cannot attend to their neighboring tokens in other windows. For example, the two tokens that contain circles cannot attend to each other, even if they are next to each other. This may cause the block-shaped artifacts, and motivates our proposed use of global attention to address this challenge.

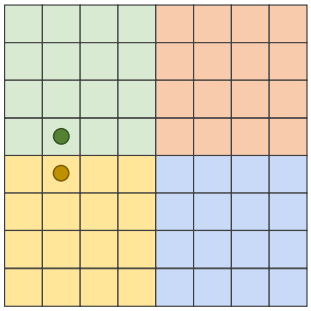

Figure 1. Illustration of the shifted window (Swin) attention which performs self-attention within local windows (window size is 4×4). Visual tokens within the same window (shown in the same color) can attend to each other, but cannot attend to their neighboring tokens in other windows. For example, the two tokens that contain circles cannot attend to each other even though they are next to each other. Hence, only using Swin attention *may* cause block-shaped artifacts (shown in Figure 7 of the main paper).

5. Additional Ablation Studies

dataset	metric	pre-training dataset		
		COCO	OI	COCO (50%)
all	PSNR↑	39.0	38.9	38.5
	MSE↓	20.9	21.5	23.7
	fPSNR↑	26.6	26.5	26.1
	fMSE↓	250.0	253.7	280.0

Table 2. Comparison of the performance of our method pre-trained using three different datasets, i.e., COCO, Open Images V6 (denoted as OI) and 50% of COCO dataset.

We investigate the influence of the pre-training dataset to the performance of our method. We present a comparison of the performance of our method pre-trained using three different datasets, *i.e.*, the unlabeled set from MS COCO (denoted as COCO), 120K images from Open Images V6 (denoted as OI) and 50% of the unlabeled set from MS COCO (denoted as 50% COCO) in Table 2.

We see that pre-training SwinIH with our method (LEMaRT) on a set of 120K randomly sampled images from Open Images V6 (roughly of same size as the unlabeled set of MS COCO). The results are comparable to those of SwinIH pre-trained on MS COCO (only 0.1 dB lower PSNR and 0.7 higher MSE on iHarmony4). This indicates that common images from the Internet can be used for LEMaRT pre-training. We see that training our model on 50% of COCO results in 0.4 dB drop in PSNR relative

to pre-training on 100% of COCO. This shows the benefit of using a larger pre-training dataset.

6. Training Time and GPU Memory Requirement.

Pre-training our model (SwinIH) does not require additional hardware and pre-training time scales linearly with dataset size. For example, pre-training SwinIH on COCO takes 54% of the time required to train/fine-tune SwinIH on iHarmony4 dataset. Compared to another Transformer-based harmonization model DHT+ [5], SwinIH uses 18% less time and 12% less GPU memory, underscoring the efficiency of our architecture.

7. User Study

We conducted a limited user study of our model compared to DHT+ [5] using real data to complement our quantitative results in the main paper. We randomly sampled 50 real composite images from RealHM. Using the method in [7], we had 7 participants who rated 1050 image pairs. The normalized B-T score for LEMaRT was 51.31, and 48.7 for DHT+, indicating our better qualitative performance. In our future work, we will conduct a more comprehensive user study with more participants and a larger number of images.

8. Acknowledgement

We sincerely thank Shixing Chen, Sheik Dawood, Chun-Hao Liu, Sajal Maheshwari and Efram Potelle for their help with the user study.

dataset	metric	composite image	DIH [13]	S^{2}AM [4]	DoveNet [3]	BargNet [1]	IntrHarm [6]	RainNet [10]	iS^{2}AM [12]	DHT+ [5]	SCS-Co [7]	LEMaRT [SwinIH]
HCOCO	PSNR↑	33.9	34.7	35.5	35.8	37.0	37.2	37.1	39.2	39.2	*39.9*	**41.0** ↑1.1
	MSE↓	69.4	51.9	41.1	36.7	24.8	24.9	29.5	16.5	15.0	*13.6*	**10.1** ↓3.5
	fPSNR↑	19.9	20.7	22.5	22.5	-	24.0	22.4	-	25.8	-	**26.9** ↑1.1
	fMSE↓	996.6	799.0	542.1	551.0	397.9	416.4	501.2	266.2	274.6	*245.5*	**209.4** ↓36.1
HAdobe	PSNR↑	8.2	32.3	33.8	34.3	35.3	35.2	36.2	38.1	37.2	*38.3*	**39.4** ↑1.1
	MSE↓	345.5	92.7	63.4	52.3	39.9	43.0	43.4	21.9	36.8	*21.0*	**18.8** ↓2.2
	fPSNR↓	17.5	22.4	24.3	25.1	-	25.9	25.0	-	*27.1*	-	**29.2** ↑2.1
	fMSE↓	2051.6	593.0	404.6	380.4	279.7	284.2	317.6	174.0	242.6	*165.5*	**147.3** ↓18.2
HFlickr	PSNR↑	28.3	29.6	30.0	30.2	31.3	31.3	31.6	33.6	33.6	*34.2*	**35.3** ↑1.1
	MSE↓	264.4	163.4	143.5	133.1	97.3	105.1	110.6	69.7	67.9	*55.8*	**40.7** ↓15.1
	fPSNR↑	18.1	19.3	20.9	20.8	-	21.6	21.0	-	23.5	-	**25.0** ↑1.5
	fMSE↓	1574.4	1099.1	785.7	827.0	698.4	716.6	688.4	443.7	471.1	*393.7*	**342.7** ↓51.0
HD2N	PSNR↑	34.0	34.6	34.5	35.3	35.7	36.0	34.8	37.7	36.4	*37.8*	**38.1** ↑0.3
	MSE↓	109.7	82.3	76.6	52.0	51.0	55.5	57.4	**40.6**	49.7	*41.8*	42.3 ↑1.7
	fPSNR↑	19.1	19.7	20.5	20.6	-	21.7	20.2	-	21.7	-	**22.8** ↑1.1
	fMSE↓	1410.0	1129.4	989.1	1075.7	835.6	797.0	916.5	*591.0*	736.6	606.8	**580.5** ↓10.5
all	PSNR↑	31.6	33.4	34.3	34.8	35.9	35.9	36.1	38.2	37.9	*38.8*	**39.8** ↑1.0
	MSE↓	172.5	76.8	59.7	52.3	37.8	38.7	40.3	24.4	27.9	*21.3*	**16.8** ↓4.5
	fPSNR↑	19.0	21.0	22.8	23.0	-	24.2	23.0	-	*25.9*	-	**27.2** ↑1.3
	fMSE↓	1376.4	773.2	594.7	532.6	405.2	400.3	469.6	265.0	295.6	*248.9*	**213.3** ↓35.6

Table 3. Our image harmonization method, LEMaRT [SwinIH], outperforms state-of-the-art (SOTA) methods on iHarmony4 across **all** four metrics including fPSNR and fMSE (due to space constraints, fPSNR and fMSE are omitted in Table 1 of the main paper). PSNR and MSE are shown in gray, as they have been shown in the Table 1 of the main paper. We repeat them for the readers' convenience. The column named *composite image* shows the results for the direct copy and paste of foreground regions on top of background images.

dataset	metric	FRIH [11]	CDTNet [2]	S^{2}CRNet [9]	Harmonizer [8]	LEMaRT [SwinIH]
HCOCO	PSNR↑	39.4	39.2	38.5	38.8	**41.0** ↑1.6
	MSE↓	15.1	16.3	23.2	17.3	**10.1** ↓5.0
HAdobe	PSNR↑	37.7	38.2	36.4	37.6	**39.4** ↑1.2
	MSE↓	23.6	20.6	34.9	21.9	**18.8** ↓1.8
HFlickr	PSNR↑	33.5	33.6	32.5	33.6	**35.3** ↑1.7
	MSE↓	68.9	68.6	98.7	64.8	**40.7** ↓27.9
HD2N	PSNR↑	37.9	38.0	36.8	37.6	**38.1** ↑0.1
	MSE↓	42.8	36.7	51.7	**33.1**	42.3 ↑8.8
all	PSNR	38.2	38.2	37.2	37.8	**39.8** ↑1.6
	MSE↓	24.0	24.7	35.6	24.3	**16.8** ↓7.2

Table 4. Comparison between our LEMaRT [SwinIH] model and four additional image harmonization models [11], [2], [9], [8] on iHarmony4. We include these models for completeness, although they underperform the state-of-the-art (SOTA) method [7] presented in the main paper.

Figure 2. Qualitative examples that show data generated for pre-training and the output of a pre-trained LEMaRT model given the generated data. We apply brightness adjustment (row 1), hue adjustment (row 2), contrast adjustment (row 3), equalization (row 4), posterization (row 5), blurring (row 6), deblurring (row 7) to generate the transformed images shown in the second column.

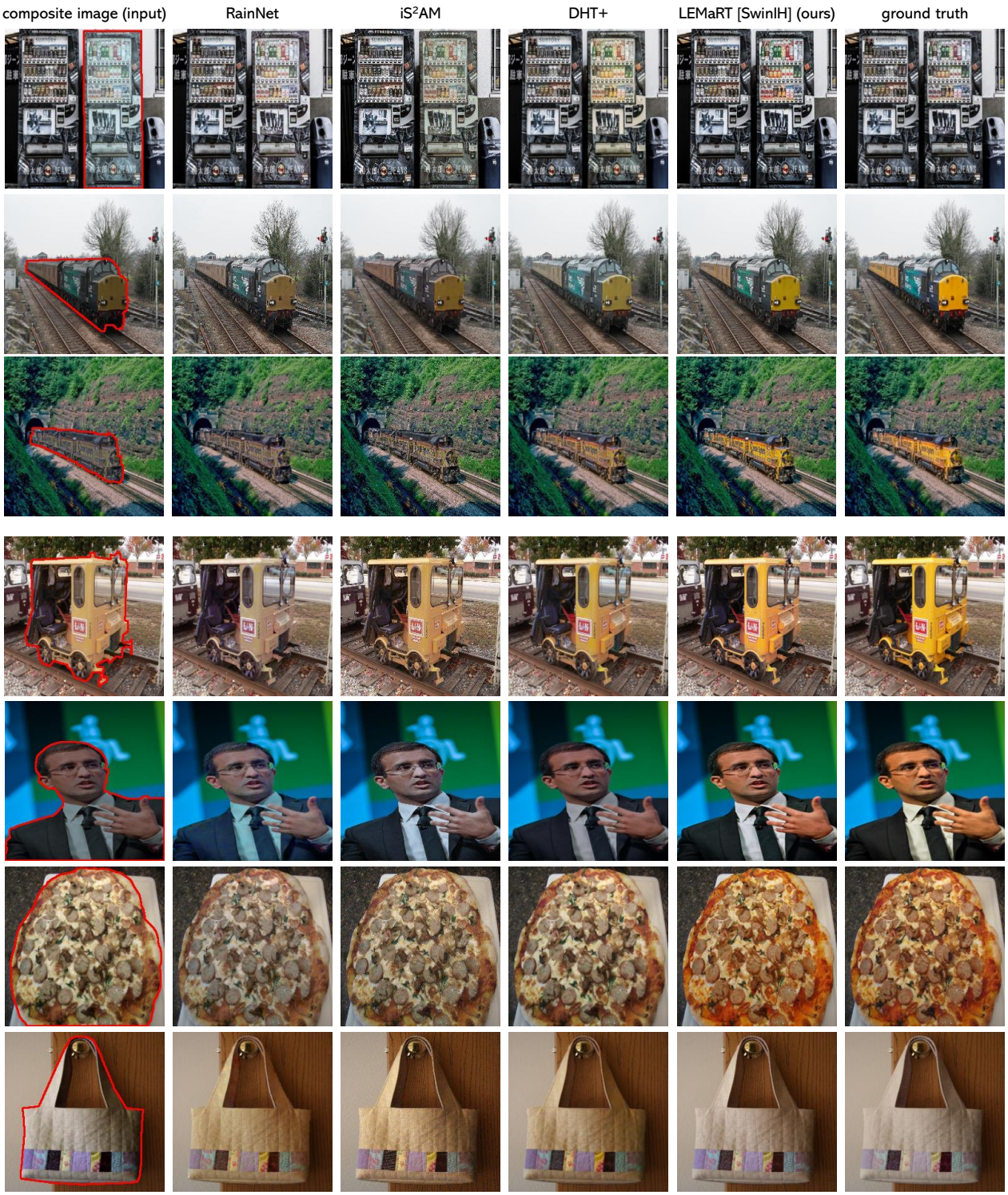

Figure 3. Qualitative comparison between our method (LEMaRT [SwinIH]) and three SOTA methods (RainNet [10], iS^{2}AM [12], DHT+ [5]) on iHarmony4. Compared to other methods, LEMaRT is better at color correction, thanks to the pre-training process during which LEMaRT learns the distribution of photo-realistic images.

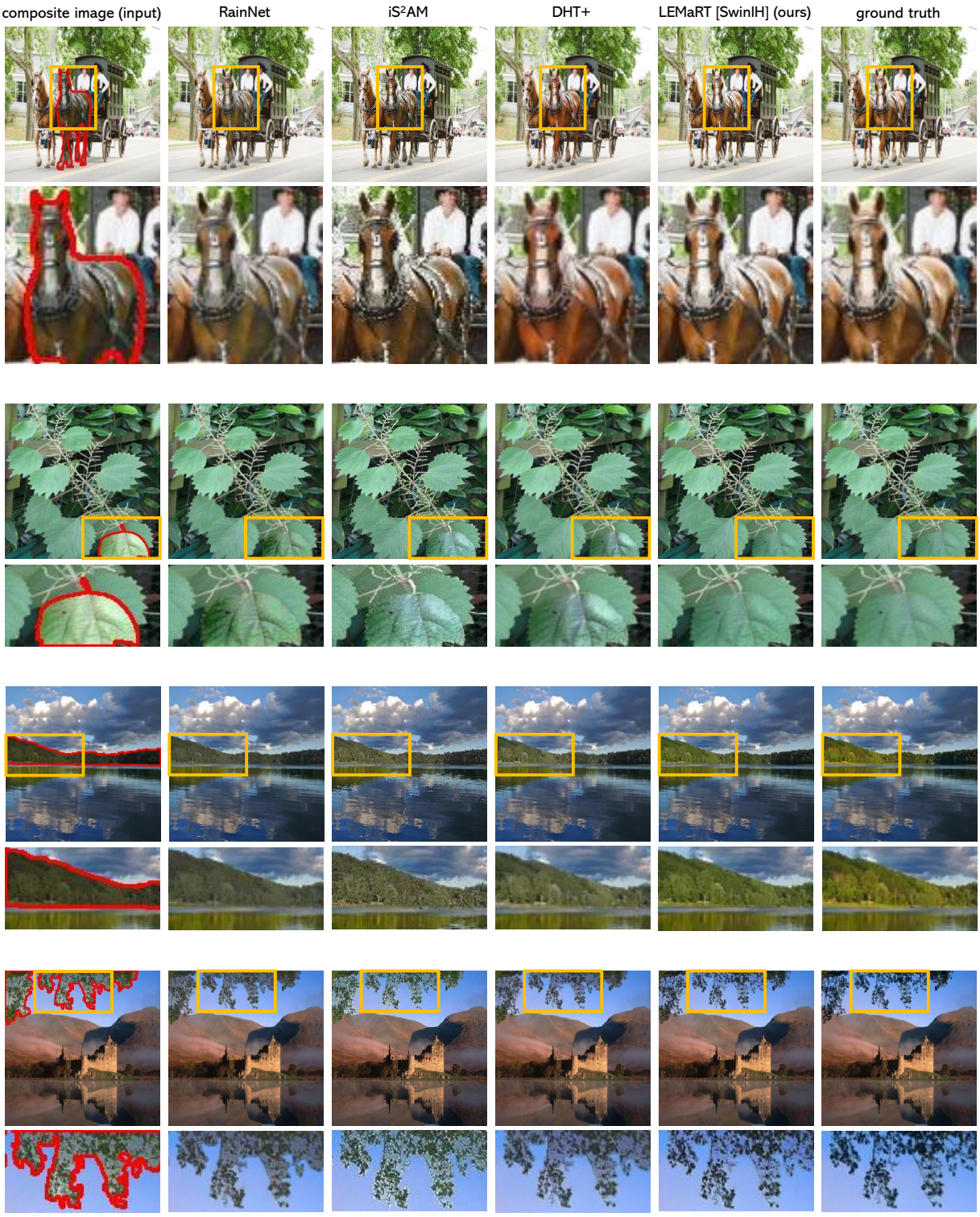

Figure 4. Qualitative comparison between our method, LEMaRT [SwinIH], and three SOTA methods (RainNet [10], iS^{2}AM [12], DHT+ [5]) on iHarmony4. We provide zoom-in views of regions in yellow rectangles. In the first example, the color of the horse in our image is more natural and closer to that of the horse in the ground truth image than other images. We see from the second example that the texture and the color of the leaf in our image are in harmony with those of the background. In the third example, the color of the mountain and its reflection are better aligned in our image than other images. We see from the forth example that our method can better harmonize subtle structures, *i.e.*, small leaves and thin branches, than other methods.

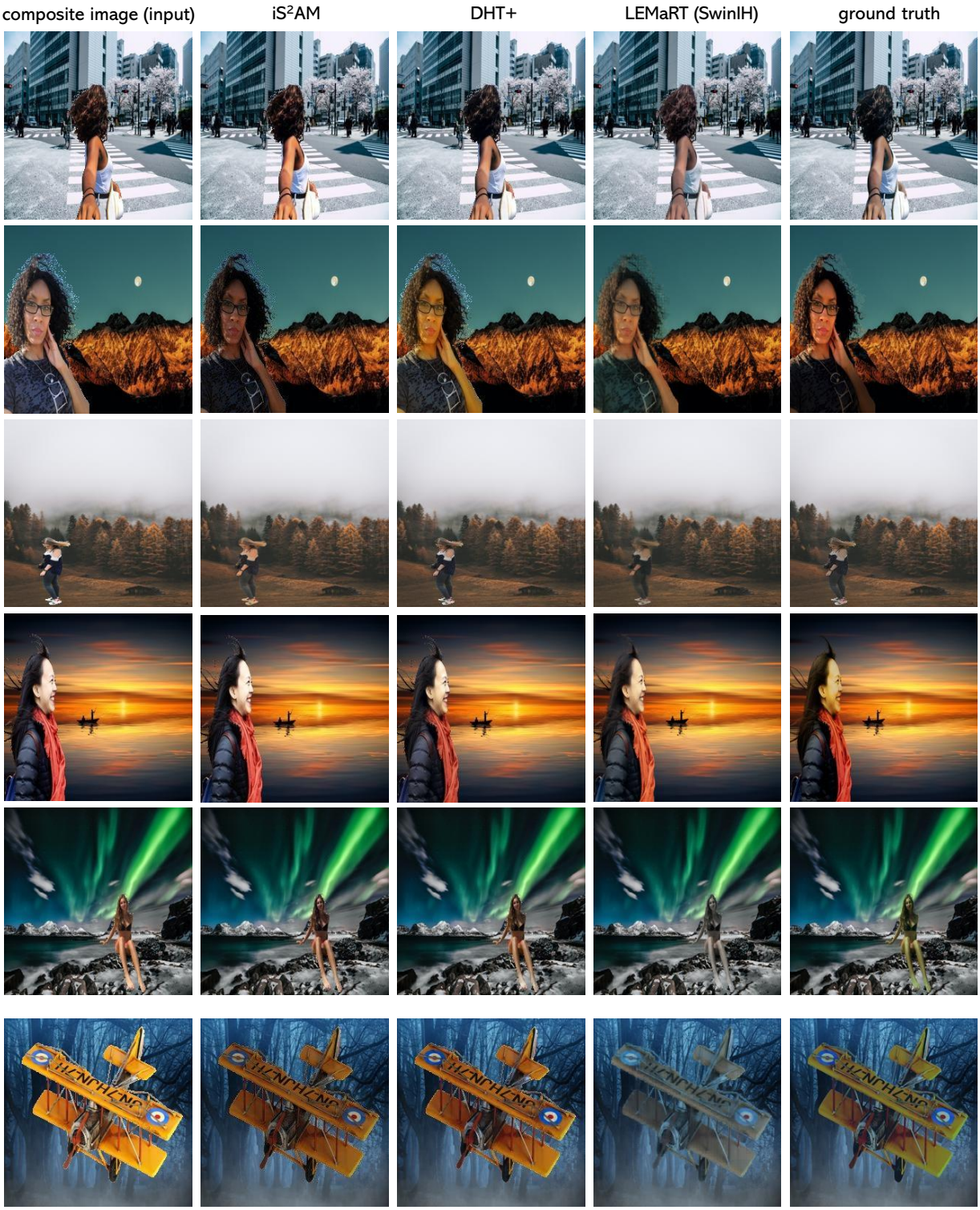

Figure 5. Qualitative comparison between our method, LEMaRT [SwinIH], and two SOTA methods (iS^{2}AM [12], DHT+ [5]) on RealHM. We see from the first five rows that our method can better harmonize a composite image than other methods. We show a controversial example in the last row. Different people may have different opinions regarding which harmonized image looks more natural.

References

[1] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Bargainnet: Background-guided domain translation for image harmonization. In *ICME*, pages 1–6, 2021.

[2] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18449–18458, 2022.

[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.

[4] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing*, 29:4759–4771, 2020.

[5] Zonghui Guo, Zhaorui Gu, Bing Zheng, Junyu Dong, and Haiyong Zheng. Transformer for image harmonization and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[6] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16367–16376, June 2021.

[7] Yucheng Hang, Bin Xia, Wenming Yang, and Qingmin Liao. Scs-co: Self-consistent style contrastive learning for image harmonization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19678–19687, 2022.

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

[9] Jingtang Liang, Xiaodong Cun, Chi-Man Pun, and Jue Wang. Spatial-separated curve rendering network for efficient and high-resolution image harmonization. In *ECCV*, 2022.

[10] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9361–9370, June 2021.

[11] Jinlong Peng, Zekun Luo, Liang Liu, Boshen Zhang, Tao Wang, Yabiao Wang, Ying Tai, Chengjie Wang, and Weiyao Lin. Frih: Fine-grained region-aware image harmonization. *arXiv preprint arXiv:2205.06448*, 2022.

[12] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *WACV*, pages 1620–1629, 2021.

[13] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, pages 3789–3797, 2017.