Spatial-Separated Curve Rendering Network for Efficient and High-Resolution Image Harmonization (Supplementary Materials)

Jingtang Liang[1*], Xiaodong Cun[2*], Chi-Man Pun[1✉], and Jue Wang[2]

[1] University of Macau, Macau, China
[2] Tencent AI Lab, Shenzhen, China
mb95464@connect.umac.mo, cmpun@umac.mo, {vinthony, arphid}@gmail.com

1 More Implementation Details

1.1 The Details of Semantic Labels Extraction

In the proposed semantic curve rendering module (SCRM), correct foreground semantic label benefits the performance of image harmonization. However, iHarmony4 [2] do not contain the ground truth labels of the foreground. To obtain the semantic labels, firstly, we get the categories in HCOCO sub-dataset via COCO API [11]. For the rest sub-datasets, we leverage a semantic segmentation model in [17] to segment the composite images. Then, we choose the segmented region which has maximal intersection with the foreground mask, and consider it as the category label. Finally, we summarize the distributions of the foreground labels of the whole dataset in Table 1. Particularly, we roughly divide the foreground regions into 5 categories, including *Person*, *Vehicle*, *Animal*, *Food* and others. We argue that this setting is also suitable for the daily usages.

Table 1: Predicted foreground distributions in iHarmony4.

Classes	HCOCO	HAdobe5k	HFlickr	Hday2night	iHarmony4
Person	13416	7274	1629	0	22319
Vehicle	4434	1338	808	10	6590
Animal	7274	747	675	0	8696
Food	6752	280	721	0	7753
Others	10952	11958	4444	434	27788
Total	42828	21597	8277	444	73146

1.2 The Details of User Study on DIH99.

As is discussed in the main paper, to evaluate the effectiveness on real-world scenarios, we conduct subjective user study to compare our proposed method

[*] These authors contribute equally to this work.
[✉] Corresponding author

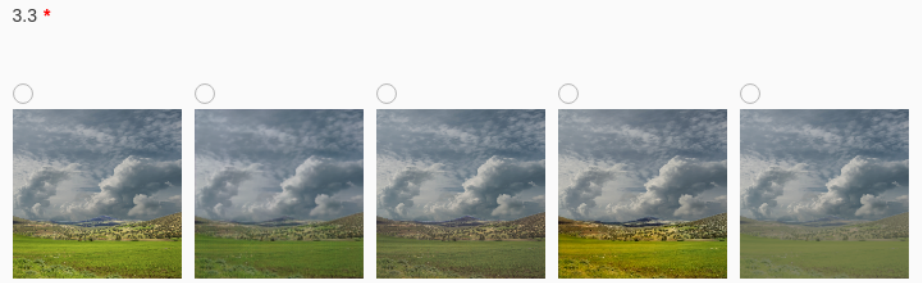


Fig. 1: The layout of a single image group example in our user study. The displaying order of the composite input and the harmonization results is randomly shuffled without annotations.

with baseline methods (DIH [14], DoveNet [2] and BargainNet [1]) on the DIH99 real composite dataset. In detail, we invite 18 participants with different ages and genders for subjective experiments. As shown in Figure 1, each participant can see a set of image groups and each group includes the original composite input and the harmonized results that generated by DIH, DoveNet, BargainNet and the proposed S^{2}CRNet. Then, we let them to select the most favorable result among different images in each image group, contributing 18×99 groups result in total. The results have been listed in Table 2 of the main paper.

2 More Experiments

2.1 Comparison with Other Similar Global Editing Methods.

To further demonstrate the effectiveness of the proposed CRM, we replace our piece-wise linear curve function in CRM by other similar global editing methods in image enhancement (3DLUT [16]) and low-light enhancement (Zero-DCE [5]), and compare the performance on HCOCO and iHarmony dataset. As summarized in Table 2, the piece-wise curve (Ours) achieves superior performance at all criteria metrics compared to other alternatives.

Table 2: Quantitative comparison in employing similar global editing methods in our CRM.

	HCOCO			iHarmony		
Methods	MSE ↓	PSNR ↑	SSIM↑	MSE↓	PSNR ↑	SSIM↑
Zero-DCE [4]	37.22	36.64	99.10	75.31	34.39	98.27
3DLUT [16]	33.22	36.95	99.18	53.05	35.49	98.77
Ours	**29.45**	**37.51**	**99.26**	**45.17**	**36.27**	**98.87**

2.2 Ablation Study for The Levels of Curve L.

In the proposed CRM, we approximate the editing curve by a L-levels piece-wise linear function. Here, we conduct the ablation experiments to investigate the influence of L by setting $L = \{32, 64, 96, 128\}$ in our S^{2}CRNet model. From Table 3, it can be inferred that approximating the curve with more levels improves the harmonizing performance. However, when L is larger than 64, increasing L has minor improvements on HCOCO and even downgrades the performance on the iHarmony dataset. It reveals that harmonizing images by a larger L will make the network hard to learn the meaningful color distribution and increase the computational cost. Hence, we set $L = 64$ in all models for a trade-off between model performance and memory computation.

Table 3: Ablation studies of the level of curve L.

Dataset	HCOCO			iHarmony		
Numer of L	PSNR↑	SSIM↑	MSE↓	PSNR↑	SSIM↑	MSE↓
32	37.60	99.24	29.13	36.17	98.83	48.31
64	**37.72**	**99.26**	**27.40**	**36.45**	**98.92**	**43.20**
96	**37.72**	**99.26**	27.81	36.24	98.88	46.63
128	37.68	**99.26**	28.58	36.19	98.86	47.00

2.3 The Effectiveness of Different Backbone.

Stronger backbone enables the networks to learn better. We evaluate the performance of different backbones in our framework as shown in Table 4. We find that more complicated structures such as VGG16 [13] perform much better than the smaller backbones, but it lacks efficiency as reported in the main paper. Also, complicated structures need more epochs to be convergent (for example, SqueezeNet-based method only needs 20 epoch to get the best result while VGG16 achieves the best performance at 48 epoch.). Thus, we report the results of best performance (VGG16 backbone) and the most efficient model (SqueezeNet backbone) in the main paper.

Table 4: Performance of different backbones in the proposed S^{2}CRNet. All experiments are trained and evaluated on iHarmony dataset under the same configurations.

Backbones	Param	PSNR ↑	SSIM ↑	MSE ↓
SqueezeNet [7]	0.95M	36.45	98.92	43.20
AlexNet [9]	2.79M	35.82	98.80	50.79
ResNet18 [6]	11.8M	36.55	98.92	41.04
VGG16 [13]	15.14M	37.18	99.01	35.58

2.4 Harmonization Performance on CPU.

Our method also shows good speed on CPU devices, which enables our method to run on the device side without any cloud computation. To this end, we compare the proposed S^{2}CRNet with other baseline methods [3,2,1,5] in harmonizing different resolution images using the same experimental environment (Intel i7-10700K CPU with 16 GB RAM on Ubuntu 18.04). Here, we choose the default SqueezeNet backbone in the proposed S^{2}CRNet for efficiency. The evaluations are conducted on the 50 images in HAdobe5k sub-dataset [2] and we present the average processing time in Table 5. The quantitative results show that our method achieves the fastest performance when operating on the CPU, and also outperforms other baselines by a large margin as the image resolution increases. Notice that our method also shows better performance than these methods as discussed in the main paper.

Table 5: Average processing time on the **CPU** under different image resolution. The best results are marked as boldface and the "NA" denotes running out of memory in our experiment.

Resolution	S^{2}AM	DoveNet	BargainNet	IIH	S^{2}CRNet
256×256	0.25s	0.05s	0.21s	1.17s	**0.03s**
512×512	0.85s	0.18s	0.75s	8.02s	**0.06s**
1024×1024	3.93s	0.79s	3.23s	NA	**0.47s**
2048×2048	NA	3.19s	13.06s	NA	**2.60s**

2.5 The Results on Different Foreground Ratios.

The differences of the composite foregrounds are also important in our task since the background is totally the same. Thus, we further compare the proposed S^{2}CRNet (including SqueezeNet [7] and VGG16 [13] backbones) with other state-of-art image harmonization approaches in different foreground ratio ranges, and the quantitative results on iHarmony4 dataset are summarized in Table 6. Following previous methods [2,1,12], we employ mean square error (MSE) and foreground mean square error (fMSE) as evaluation metrics, where fMSE measures the MSE scores of the harmonized foreground regions. We follow previous works [2,1] to evaluate the performance in four different foreground ranges, including 0% to 5%, 5% to 15%, 15% to 100% and overall results. As shown in Table 6, the performance of all the models will be downgraded as the foreground ratios increase. Nevertheless, our S^{2}CRNet-SqueezeNet achieves the best performance in most of the foreground ratio intervals especially on small foreground regions (0%-5% and 5%-15% foreground ratios). Furthermore, when employing VGG16 backbone (S^{2}CRNet-VGG16), our method achieves state-of-art performance and outperforms other methods by a large margin in all the foreground ratio intervals.

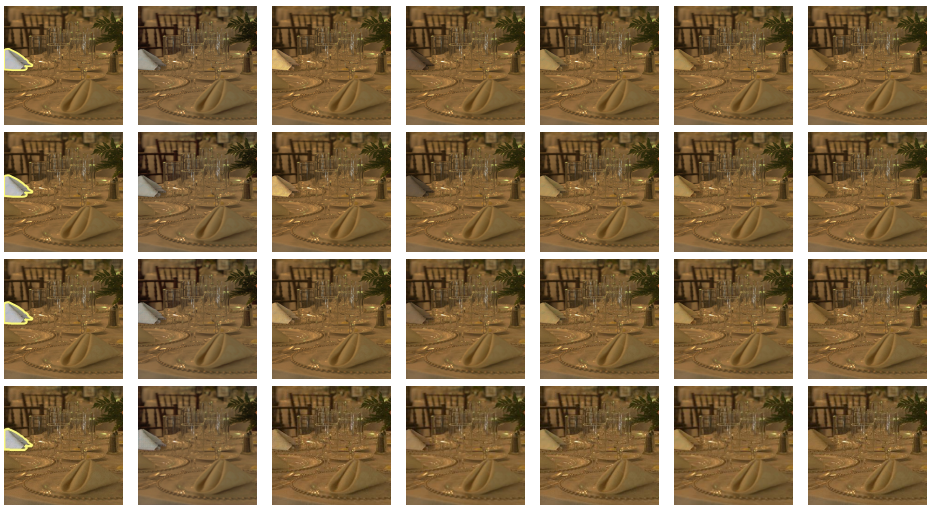


Fig. 2: Qualitative comparisons with existing methods in harmonizing images at different resolutions.From left to right are (a)Input (b) DoveNet [2], (c) BargainNet [1], (d) S^{2}AM [3], (e) S^{2}CRNet-S (Ours), (f) S^{2}CRNet-V (Ours) and (g) Target. Here, we resize the all the images to the same resolutions for presentation. The original input images are 256×256, 512×512, 1024×1024, 2048×2048 from the bottom up successively. We mark the composite foreground mask as yellow region. S^{2}CRNet-S and S^{2}CRNet-V denote our method employing SqueezeNet and VGG16 backbone, respectively.

2.6 Visual Comparison on High-Resolution Images

Our method shows the resolution-invariant results that benefits from the proposed curve-based framework. Here, we visualize an example to show the influence of the input resolution in different methods. Similar to the high-resolution image harmonization experiments in the primary paper, we compare our method with other baseline methods [2,1,3] in harmonizing images at different resolutions including the square of 256, 512, 1024 and 2048. As shown in Figure 2, due to the changes of reception fields, the other state-of-the-art methods show unstable results. Differently, both the proposed S^{2}CRNet-SqueezeNet and S^{2}CRNet-VGG16 get more stable and favorable results, while the others show downgraded harmonization qualities as the resolutions increase.

2.7 More Rendering Curves Visualization

We present more visual results to visualize the preliminary rendering curves and the curves in cascaded refinements. As shown in Figure 3, the curves generated by the CRM are different according to various input samples, which demonstrates

Table 6: Foreground Harmonization Comparisons on iHarmony4. The fMSE measures the mean square error scores of the harmonized foreground regions. The best and the second best are marked as boldface and underline respectively.

Foreground Ratios	0%-5%		5%-15%		15%-100%		0%-100%	
Evaluation metric	MSE ↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓
Input composition	28.51	1208.86	119.19	1323.23	577.58	1887.05	172.47	1387.30
Xue *etal.* [15]	41.52	1481.59	120.62	1309.79	444.65	1467.98	150.53	1433.21
Lalonde & Efros [10]	31.24	1325.96	132.12	1459.28	479.53	1555.69	155.87	1411.40
Zhu *etal.* [18]	33.30	1297.65	145.14	1577.70	682.69	2251.76	204.77	1580.17
DIH[14]	18.92	799.17	64.23	725.86	228.86	768.89	76.77	773.18
DoveNet[2]	14.03	591.88	44.90	504.42	152.07	505.82	52.36	549.96
S^2AM[3]	13.51	509.41	41.79	454.21	137.12	449.81	48.00	481.79
BargainNet[1]	10.55	450.33	32.13	359.49	<u>109.23</u>	<u>353.84</u>	<u>37.82</u>	405.23
IIH[5]	9.97	441.02	31.51	363.61	110.22	354.84	38.71	400.29
RainNet[12]	11.66	550.38	32.05	378.69	117.41	389.81	40.29	469.61
S^2CRNet-SqueezeNet	<u>8.42</u>	<u>301.97</u>	<u>29.74</u>	<u>336.24</u>	126.56	405.13	43.21	<u>336.99</u>
S^2CRNet-VGG16	**6.80**	**239.94**	**25.37**	**271.70**	**103.42**	**333.96**	**35.58**	**274.99**

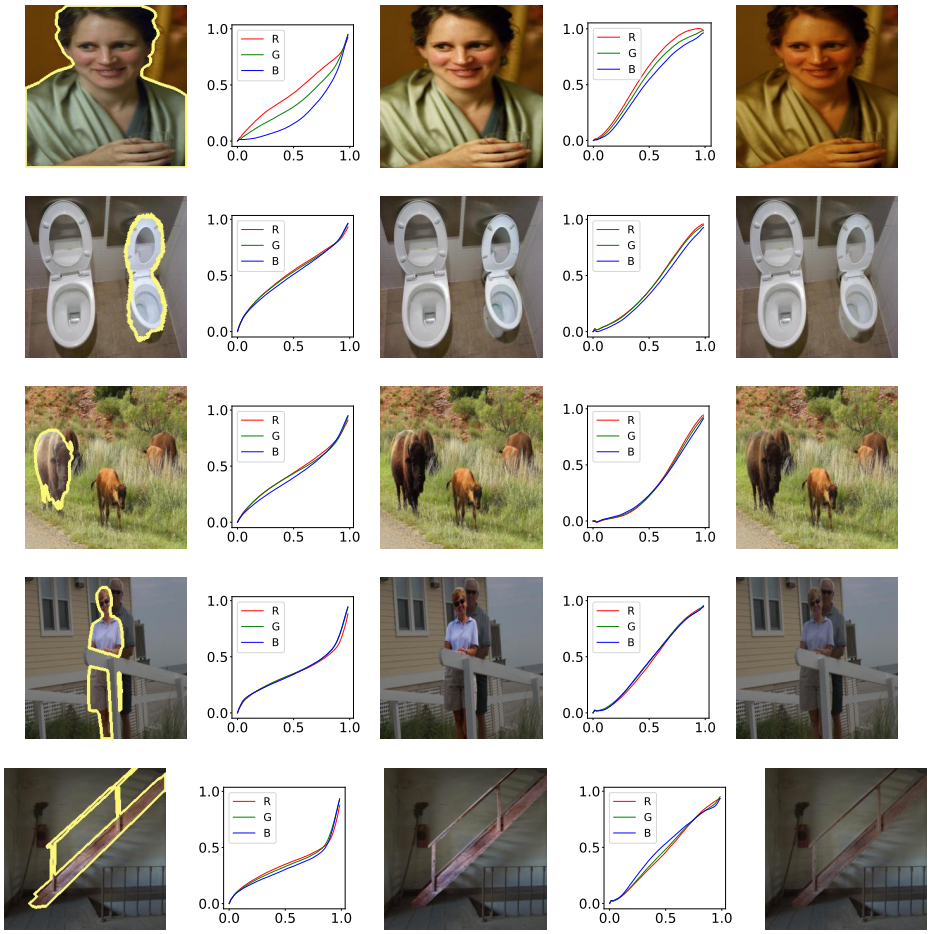
that the S^2CRNet can produce the practical curve parameters for each images via the deep features. Also, for cascaded refinement, the curves in each stage are also different and these stage-aware curves contribute to the improvement of the harmonization performance according to the visualized results of different stages in Figure 3.

2.8 Visualized Comparison in DIH99 and RealHM Dataset

To demonstrate the effectiveness of the proposed method on real-world scenarios, we further evaluate the proposed S^2CRNet with two backbones (SqueezeNet [7] and VGG16 [13]) and the baseline methods (DIH [14], DoveNet [2] and BargainNet [1]) on DIH99 [14] and RealHM [8] real composite dataset, and visualize the harmonization results in Figure 4 and Figure 5. As shown in Figure 4 and and Figure 5, the proposed efficient S^2CRNet can also achieve favorable results on real composite images compared to other presented methods, showing the reliable generalization in real-scenario applications.

2.9 More Visual Results on iHarmony4 Dataset

Given some composite images and their foreground masks, in Figure 6, we present more harmonized results generated by methods including S^2AM [3], DoveNet [2], BargainNet [1] and our S^2CRNet on iHarmony4 dataset. Compared with the other baselines, both S^2CRNet-SqueezeNet and S^2CRNet-VGG16 can generate more harmonious results and also maintain visual similarities with the target natural images.



(a) Input (b) 1 st Curve (c) 1 st Curve (d) 1 st Curve (e) 1 st Curve

Fig. 3: More visualized results of the cascaded rendering curves generated by S^{2}CRNet. We mark the composite foreground mask as yellow region.

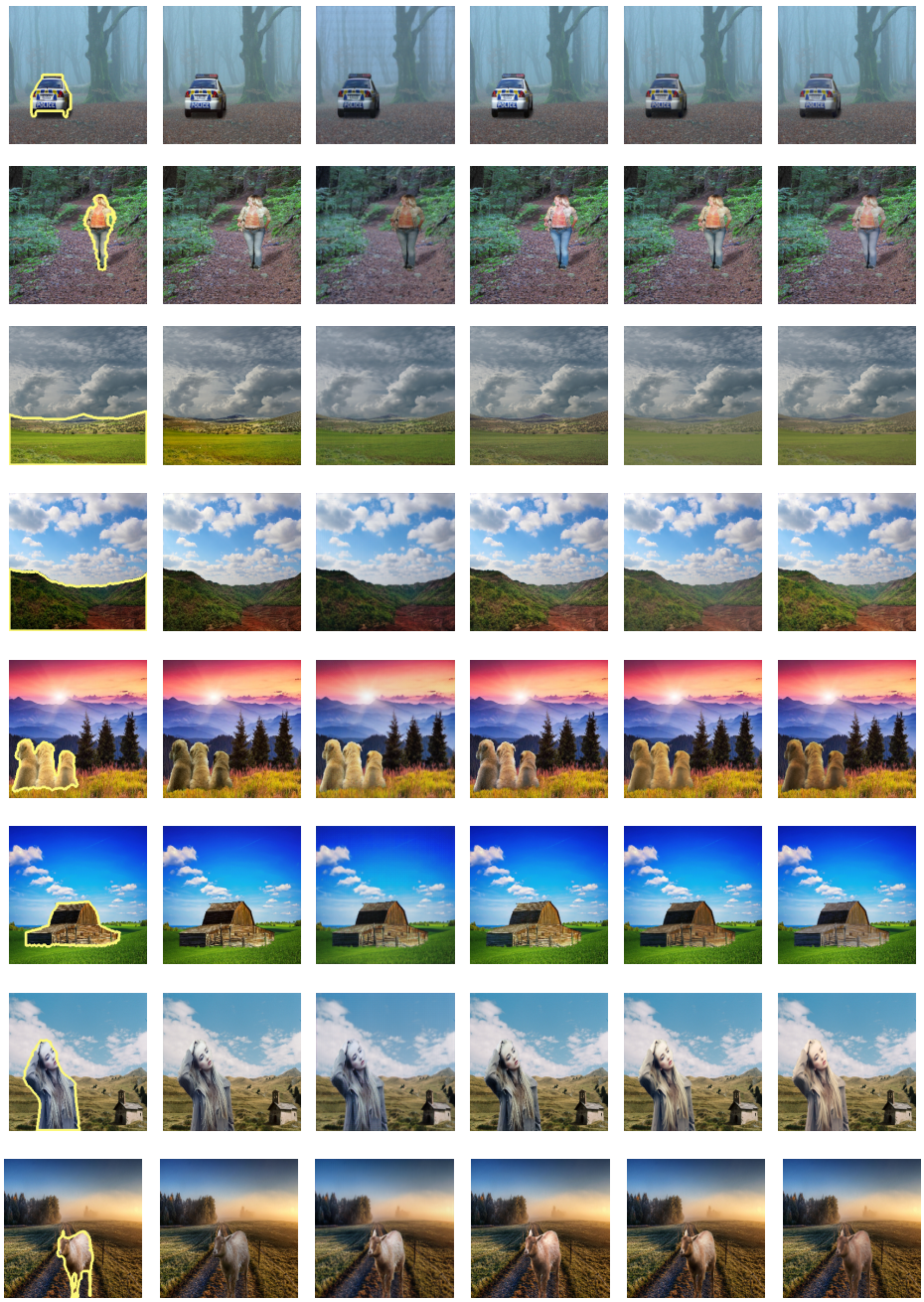
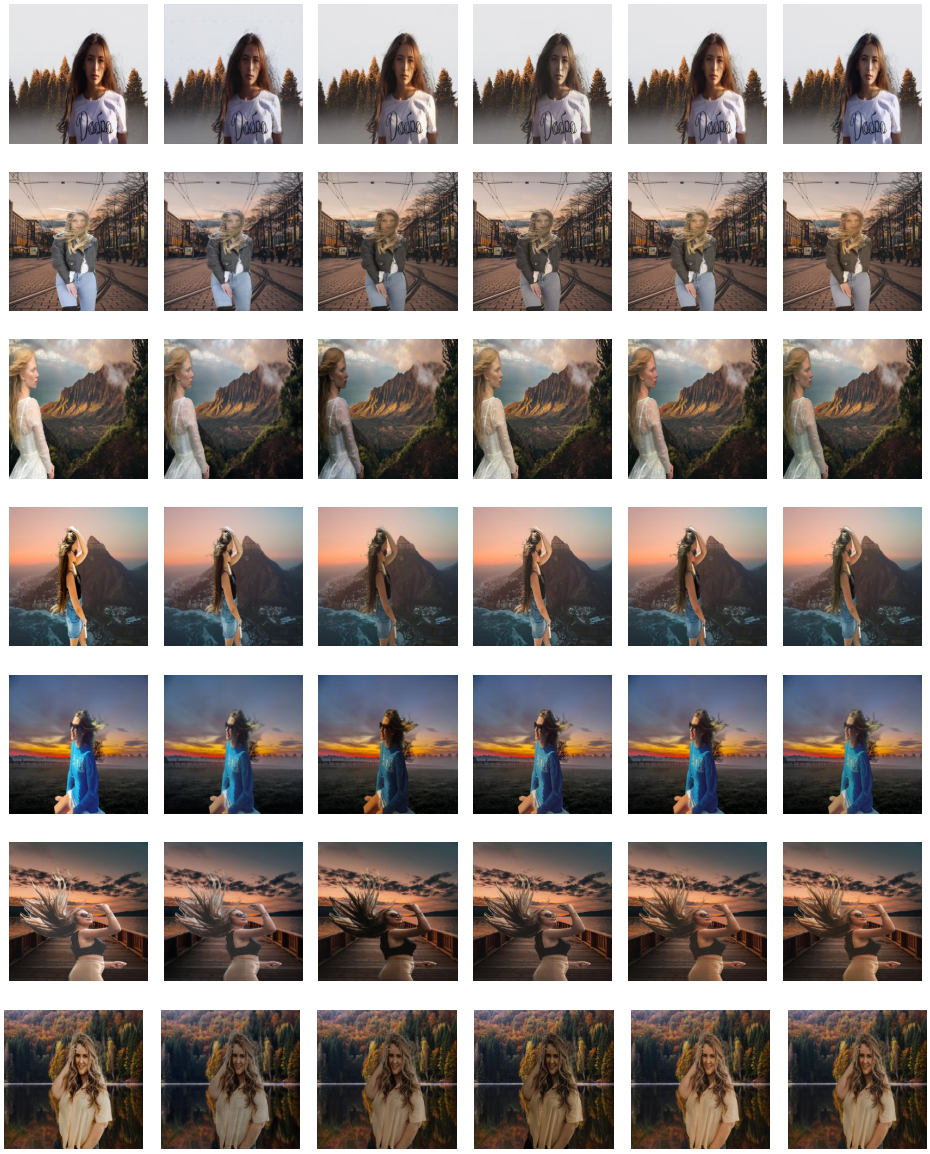


Fig. 4: More comparisons with baseline methods [14,2,1] on DIH99 dataset. From left to right are (a)Input (b) DIH [14], (c) DoveNet [2], (d) BarGainNet [1], (e) S^2CRNet-S (Ours) and (f) S^2CRNet-V (Ours). We mark the composite foreground mask as yellow region. S^2CRNet-S and S^2CRNet-V denote our method employing SqueezeNet and VGG16 backbone, respectively.



(a) (b) (c) (d) (e) (f)

Fig. 5: More comparisons with baseline methods [3,2,8] on RealHM dataset [8]. From left to right are (a)Input (b) DoveNet [2],(c) S^{2}AM [3], (d) SSH [8], (e) S^{2}CRNet-S (Ours) and (f) S^{2}CRNet-V (Ours). S^{2}CRNet-S and S^{2}CRNet-V denote our method employing SqueezeNet and VGG16 backbone, respectively.

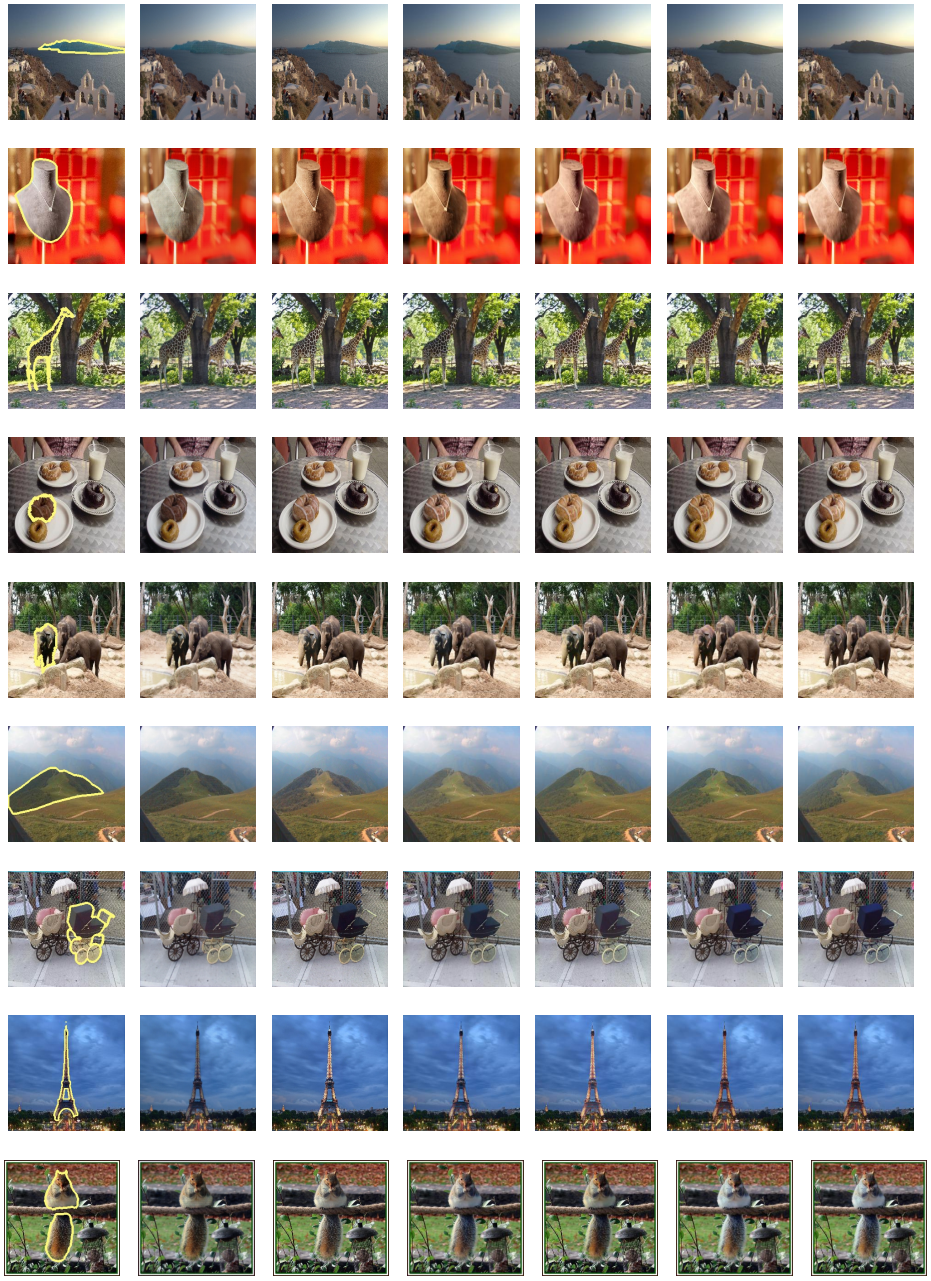


Fig. 6: More qualitative comparison with other methods [2,1,3] on iHarmony4 Dataset. From left to right are (a)Input (b) DoveNet [2], (c) BarGainNet [1], (d) S^2AM [3], (e) S^2CRNet-S (Ours), (f) S^2CRNet-V (Ours) and (g) Target. We mark the composite foreground mask as yellow region. S^2CRNet-S and S^2CRNet-V denote our method employing SqueezeNet and VGG16 backbone, respectively.

References

1. Cong, W., Niu, L., Zhang, J., Liang, J., Zhang, L.: BargainNet: Background-guided domain translation for image harmonization. In: ICME (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. Cun, X., Pun, C.M.: Improving the harmony of the composite image by spatial-separated attention module. TIP **29**, 4759–4771 (2020)
4. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: CVPR. pp. 1780–1789 (2020)
5. Guo, Z., Zheng, H., Jiang, Y., Gu, Z., Zheng, B.: Intrinsic image harmonization. In: CVPR. pp. 16367–16376 (June 2021)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
7. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)
8. Jiang, Y., Zhang, H., Zhang, J., Wang, Y., Lin, Z., Sunkavalli, K., Chen, S., Amirghodsi, S., Kong, S., Wang, Z.: Ssh: A self-supervised framework for image harmonization. In: ICCV. pp. 4832–4841 (2021)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NIPS **25**, 1097–1105 (2012)
10. Lalonde, J.F., Efros, A.A.: Using color compatibility for assessing image realism. In: ICCV. pp. 1–8. IEEE (2007)
11. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
12. Ling, J., Xue, H., Song, L., Xie, R., Gu, X.: Region-aware adaptive instance normalization for image harmonization. In: CVPR. pp. 9361–9370 (June 2021)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
14. Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.H.: Deep image harmonization. In: CVPR (2017)
15. Xue, S., Agarwala, A., Dorsey, J., Rushmeier, H.: Understanding and improving the realism of image composites. TOG **31**(4), 1–10 (2012)
16. Zeng, H., Cai, J., Li, L., Cao, Z., Zhang, L.: Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. TPAMI (2020)
17. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. IJCV (2018)
18. Zhu, J.Y., Krahenbuhl, P., Shechtman, E., Efros, A.A.: Learning a discriminative model for the perception of realism in composite images. In: ICCV. pp. 3943–3951 (2015)