Supplementary Material for High-Resolution Image Harmonization via Collaborative Dual Transformations

Wenyan Cong[1], Xinhao Tao[2], Li Niu[1]*, Jing Liang[1], Xuesong Gao[3,4], Qihao Sun[4], Liqing Zhang[1]
[1] Shanghai Jiao Tong University [2] Harbin Institute of Technology [3] Tianjin University [4] Hisense
[1]{plcwyam17320,ustcnewly,leungjing}@sjtu.edu.cn [2]1180300213@stu.hit.edu.cn
[3,4]gaoxuesong@tju.edu.cn [4]sunqihao@hisense.com [1]zhang-lq@cs.sjtu.edu.cn

In this supplementary, we will first introduce how we transplant two baseline groups to high-resolution image harmonization task in Section 1. Then, we will investigate the performance on different foreground ratio ranges in Section 2. We will conduct ablation studies to analyze the roles of three components in our CDTNet in Section 3, including both qualitative analysis in Section 3.1 and efficiency analysis in Section 3.2. Then, we will take an in-depth look at our color mapping module in Section 4. Besides, we will provide more visual results of different methods on two different resolutions (*i.e.*, 1024×1024 and 2048×2048) in Section 5. Furthermore, we will introduce the details of our created 100 high-resolution real composite images and the conducted user study, and exhibit some harmonized results of different methods on real composite images in Section 6. Finally, we will discuss the limitations of our method in Section 7. *Note that when conducting experiments on* 1024×1024 *(resp.,* 2048×2048*) resolution, the resolution of low-resolution generator in our CDTNet is set as 256 (resp., 512), unless otherwise stated.*

1. Baseline Transplantation

Since there are no existing high-resolution image harmonization methods available for comparison, we transplant low-resolution image harmonization methods [4–6, 9, 10] and high-resolution image-to-image translation methods [1, 3, 12] to our task with essential modification of their official implementation. The low-resolution image harmonization models [4–6, 9, 10] can be trained on high-resolution images despite the huge memory consumption. Thus, we train these models on high-resolution images with sufficient GPU memory. For high-resolution image-to-image translation methods, we modify their input by concatenating foreground mask and the composite image, leaving the other components of the network untouched, because the foreground mask has been proved essential for the harmonization task [4, 5].

*Corresponding author.

2. Foreground Ratio Ranges Analyses

To better demonstrate that high-resolution pixel-to-pixel transformation may be weak in capturing long-range dependency due to local convolution operations [13], we investigate the performance of iS^{2}AM and our CDTNet in different foreground ratio ranges based on MSE and foreground MSE (fMSE) metrics. The results are reported in Table1. On 1024×1024 resolution, iS^{2}AM achieves a comparable performance to our CDTNet when the foreground ratios are less than 15%. However, when the foreground ratios are greater than 15%, our CDTNet outperforms iS^{2}AM by a large margin. On 2048×2048 resolution, our CDTNet outperforms iS^{2}AM in all ranges of foreground ratios. Moreover, the performance gap increases as the foreground ratio range increases, which strongly demonstrates that iS^{2}AM tends to have inferior performance especially when the foregrounds are large.

3. Ablation Studies

3.1. Qualitative Analyses

Our CDTNet consists of a low-resolution generator for pixel-to-pixel transformation, a color mapping module for RGB-to-RGB transformation, and a refinement module to take advantage of both. We have provided the quantitative results of ablating each component in Table 4 in the main paper. In Figure 1, we present some example images harmonized by different ablated versions on 1024×1024 resolution, including only using the low-resolution generator (row 1 in Table 4 in the main paper), only using the color mapping module (row 2 in Table 4 in the main paper), and our full method (row 10 in Table 4 in the main paper). We can observe that the upsampled results of the low-resolution generator are too blurry to meet the satisfaction of high-resolution image harmonization. In contrast, the results of the color mapping module are with high resolution and sharp contour. However, due to the lack of fine-grained information, global RGB-to-RGB transformation

Image Size	Foreground ratios	$0\% \sim 5\%$		$5\% \sim 15\%$		$15\% \sim 100\%$		$0\% \sim 100\%$	
		MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓
1024 ×1024	Input composite	47.70	1810.54	179.51	1922.07	827.27	2630.86	352.05	2122.37
	iS^{2}AM [10]	4.91	192.56	13.98	151.67	56.19	162.23	25.03	168.85
	CDTNet	**4.34**	**178.50**	**13.32**	**146.95**	**44.80**	**132.20**	**21.24**	**152.13**
2048 ×2048	Input composite	48.28	1834.87	180.91	1938.20	830.90	2642.80	353.92	2139.97
	iS^{2}AM [10]	6.51	261.54	23.47	247.15	108.99	305.52	46.37	271.59
	CDTNet	**3.98**	**161.13**	**12.67**	**140.14**	**51.14**	**145.64**	**23.35**	**159.13**

Table 1. MSE and foreground MSE (fMSE) of iS^{2}AM and our CDTNet in each foreground ratio range based on the whole test set. The best results are denoted in boldface.

cannot leverage local context and may produce unsatisfactory local results (*e.g.*, darker roofs in row 2, brighter signs and words in row 4, more obvious reflective light in row 6). Our full method takes advantage of both modules and produces more favorable results, which are visually pleasant and closer to the ground-truth.

3.2. Efficiency Analyses

In Table 3 in the main paper, we evaluate the efficiency of our method and its simplified variant. Here, we further investigate the efficiency for each individual module (*i.e.*, low-resolution generator **G**, color mapping module **C**, and refinement module **R**) in Table 2.

Since the generator only operates on low-resolution inputs, its time consumption, as well as computational and memory cost, are well-surpressed compared to that of iS^{2}AM [10] in Table 3 in the main text. Note that the color mapping module in Table 2 does not include the encoder E from the low-resolution generator. Since global RGB-to-RGB transformation is barely constrained by the number of pixels and the weight predictor is light-weighted in structure, our color mapping module is very efficient in high resolutions. Our refinement module is also simple in structure. Although its memory cost and FLOPs are relatively higher, the inference time is fast, and the resolution change only causes a slight increase in the time cost.

Image Size	Module	Time↓	Memory↓	FLOPs↓
1024 ×1024	**G**	8.6	322	14.96
	C	1.1	49	0.07
	R	1.1	1552	63.02
2048 ×2048	**G**	9.2	1287	59.84
	C	1.1	196	0.27
	R	1.2	6208	252.07

Table 2. Efficiency evaluation for each individual module of our CDTNet on 1024×1024 and 2048×2048 resolutions, including "Time" (ms), "Memory" (MB), and "FLOPs" (G). **G**: low-resolution generator. **C**: color mapping module. **R**: refinement module.

4. Analyses of Our Color Mapping Module

In our color mapping module, we employ N basis transformations (LUTs) and a weight predictor to predict the combination coefficients of basis transformations. In the main paper, we set $N = 4$ by default. In this section, we first investigate the number N of the basis transformations on 1024×1024 resolution. Specifically, we set $N = \{1, 2, 3, 4, 5, 6, 7, 8\}$ and employ weight predictor as proposed in Section 3.2 in the main text.

In Figure 2, we plot the performance by varying N. It can be seen that using only a single LUT has poor performance due to the weak transformation ability. When increasing N from 1 to 4, the performance is boosted obviously because more LUTs can improve the expressiveness of image-specific color transformation. Further increasing N from 4 to 8 leads to the performance convergence with only minor improvements. Since more LUTs will increase the memory consumption, we set $N = 4$ by default in all experiments for a good trade-off between performance and memory consumption.

We also present example images harmonized using different numbers N of LUTs, where N is set as $\{1, 2, 4\}$ and the results are shown in Figure 3. It can be seen that as N increases, the harmonized results become more visually appealing and closer to the ground-truth image, because more basis LUTs make the combined transformation more expressive.

We also take an in-depth look at the learnt N LUTs and observe the transformed result using each LUT. One interesting observation is that when $N > 1$, the transformed result using the first LUT is close to the composite image while the transformed results using the other $N-1$ LUTs look like residues. This might be caused by our way of initializing N LUTs. Specifically, the first LUT is initialized as an identity map while the other $N-1$ LUTs are initialized as zero maps. Therefore, the combination of transformed results using N LUTs is equivalent to making adjustments for the composite image by adding proper residues.

Figure 1. Example results harmonized by only using the low-resolution generator $\hat{\mathbf{I}}^{lr}_{pix}$ (row 1 in Table 4 in the main paper), only using the color mapping module $\hat{\mathbf{I}}^{hr}_{rgb}$ (row 2), and our full method CDTNet (row 10). The red border lines indicate the foreground, and the yellow boxes zoom in the particular regions for a better observation.

5. More Visualization Results on HAdobe5k

We provide more results of baseline pix2pixHD [12], iS^{2}AM [10], our simplified variant which uses deep RGB-to-RGB transformation only (CDTNet-256 (sim) in Table 1 in the main text), and our CDTNet on 1024×1024 resolution in Figure 4. We also provide additional results of iS^{2}AM, our simplified variant (CDTNet-512 (sim) in Table 1 in the main text), and our CDTNet on 2048×2048

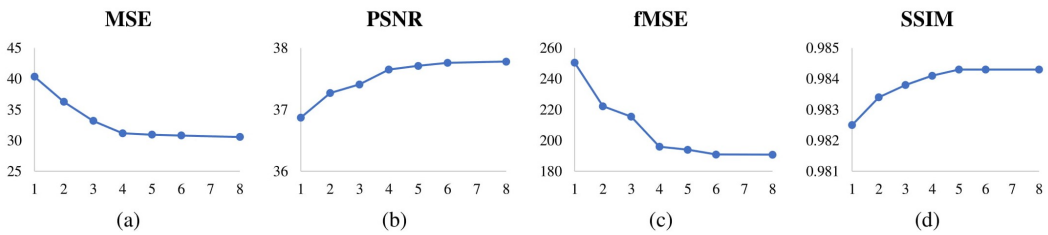

Figure 2. Impact of the number N of the basis transformations on 1024×1024 resolution.

resolution in Figure 5. pix2pixHD [12] is not specifically designed for image harmonization, so its performance is less satisfactory. The large image harmonization models directly trained on high-resolution images may be weak in capturing long-range dependency, as discussed in Section 2. In Figure 4, our simplified variant (CDTNet (sim)) obtains globally reasonable illumination but insufficient local harmony, while in Figure 5, CDTNet (sim) outperforms iS^{2}AM by generating more harmonious results, which demonstrates the expressiveness of our deep RGB-to-RGB transformation. Based on Figure 4 and Figure 5, for both resolutions, our CDTNet could generate more plausible and satisfactory harmonization results stably and adaptively, which demonstrates the superiority and robustness of our method.

6. Results on High-Resolution Real Composite Images

Considering that the composite images in HAdobe5k are synthesized composite images, we further perform evaluation on 100 high-resolution real composite images.

6.1. Image Statistics

We create high-resolution real composite images using the images from Open Image Dataset V6 [7]. Open Image Dataset V6 contains ∼9M images with 28M instance segmentation annotations of 350 categories, where enormous images are collected from Flickr [1] and with high resolution. Therefore, we collect the foreground images from the whole Open Image Dataset V6 and use the provided instance segmentation masks to crop the foregrounds. To ensure the diversity and quality of the composite images, we collect the background images from both Open Image Dataset V6 and Flickr, considering the resolutions and semantics. Then, we use PhotoShop to combine cropped foregrounds and background images by placing the foreground region at a reasonable location with a suitable scale. After that, we choose 100 high-resolution real composite images with obviously inharmonious foreground and background for evaluation.

The generated real composite images are with random resolution from 1024 to 6016. The foregrounds include human portraits and general objects (*e.g.*, dog, cat, car), and the backgrounds cover diverse scenes. Example high-resolution real composite images and corresponding masks could be found in Figure 6.

[1] https://www.flickr.com

Method	B-T score↑
Composite	0.999
pix2pixHD [12]	0.386
iS^{2}AM [10]	1.076
CDTNet	1.216

Table 3. B-T scores of baseline pix2pixHD [12], iS^{2}AM [10], and our CDTNet on 100 high-resolution real composite images.

6.2. User Study

To demonstrate the effectiveness of our proposed CDTNet in real scenarios, we follow [4, 5, 11] and further compare our model with baseline pix2pixHD [12], iS^{2}AM [10] on 100 real composite images resized to 1024×1024 resolution. More specifically, given each composite image and its 3 harmonized outputs from 3 different methods, we can construct image pairs (I_i, I_j) by randomly selecting two from these 4 images $\{I_i|_{i=1}^4\}$. Hence, we can construct 600 image pairs based on 100 real composite images.

Each user involved in this subjective evaluation could see an image pair each time to decide which one looks more harmonious and realistic. Considering the user bias, 14 users participate in the study in total, contributing 8400 pairwise results. With all pairwise results, we employ the Bradley-Terry (B-T) model [2, 8] to obtain the global ranking of all methods, and the results are reported in Table 3. Our proposed method shows an advantage over other methods with the highest B-T score, which demonstrates that by combining the complementary pixel-to-pixel transformation and RGB-to-RGB transformation, our method could generate more favorable results in real-world applications.

6.3. Qualitative Results

To visualize the comparison on high-resolution real composite images, we provide the harmonization results of pix2pixHD [12], iS^{2}AM [10], and our CDTNet in Figure

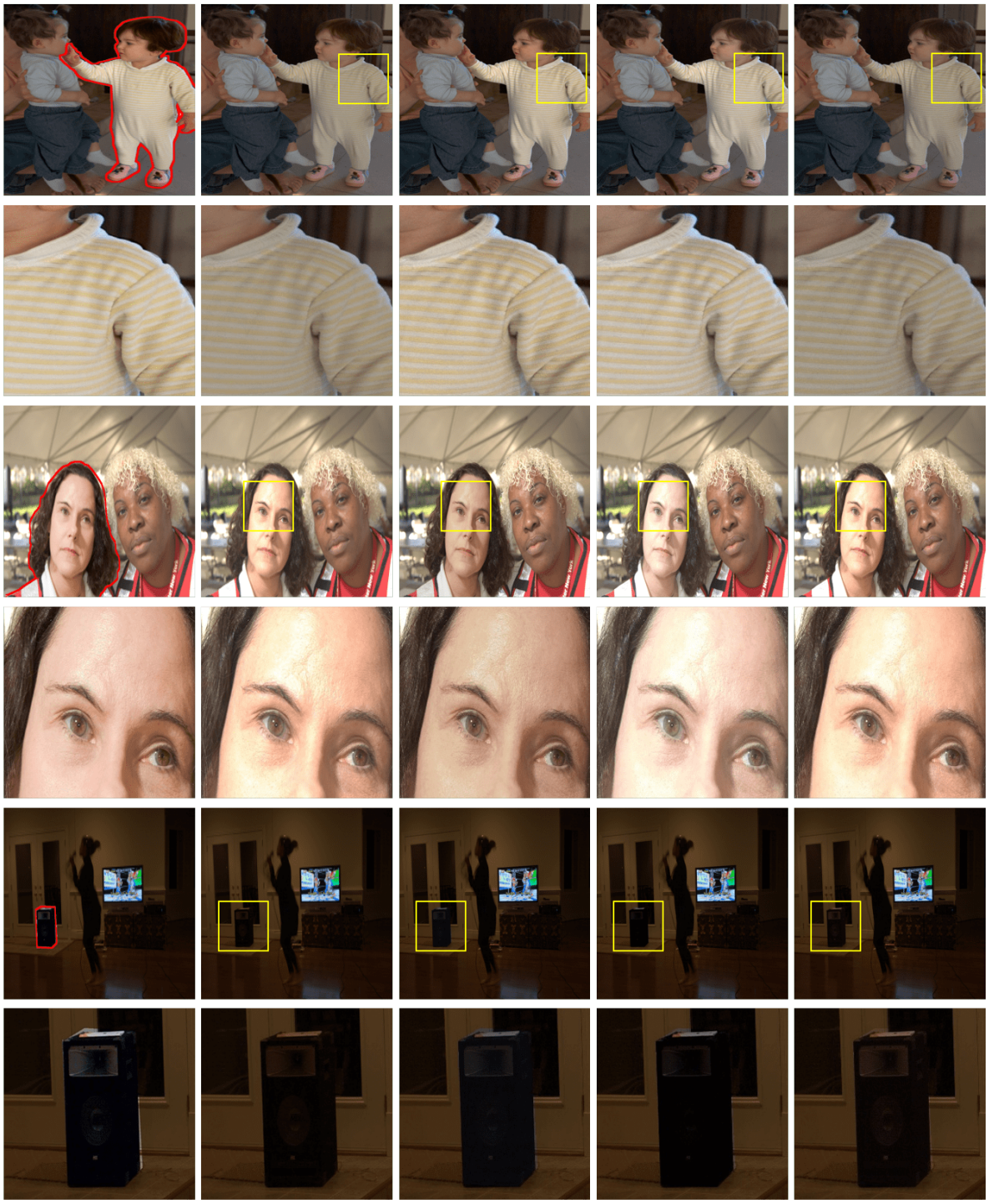

Figure 3. Example harmonized results using different numbers of LUTs in the color mapping module. From left to right, we show the input composite image, the ground-truth image, as well as the harmonized results generated using 1 LUT, 2 LUTs, and 4 LUTs (our setting) on 1024×1024 resolution.

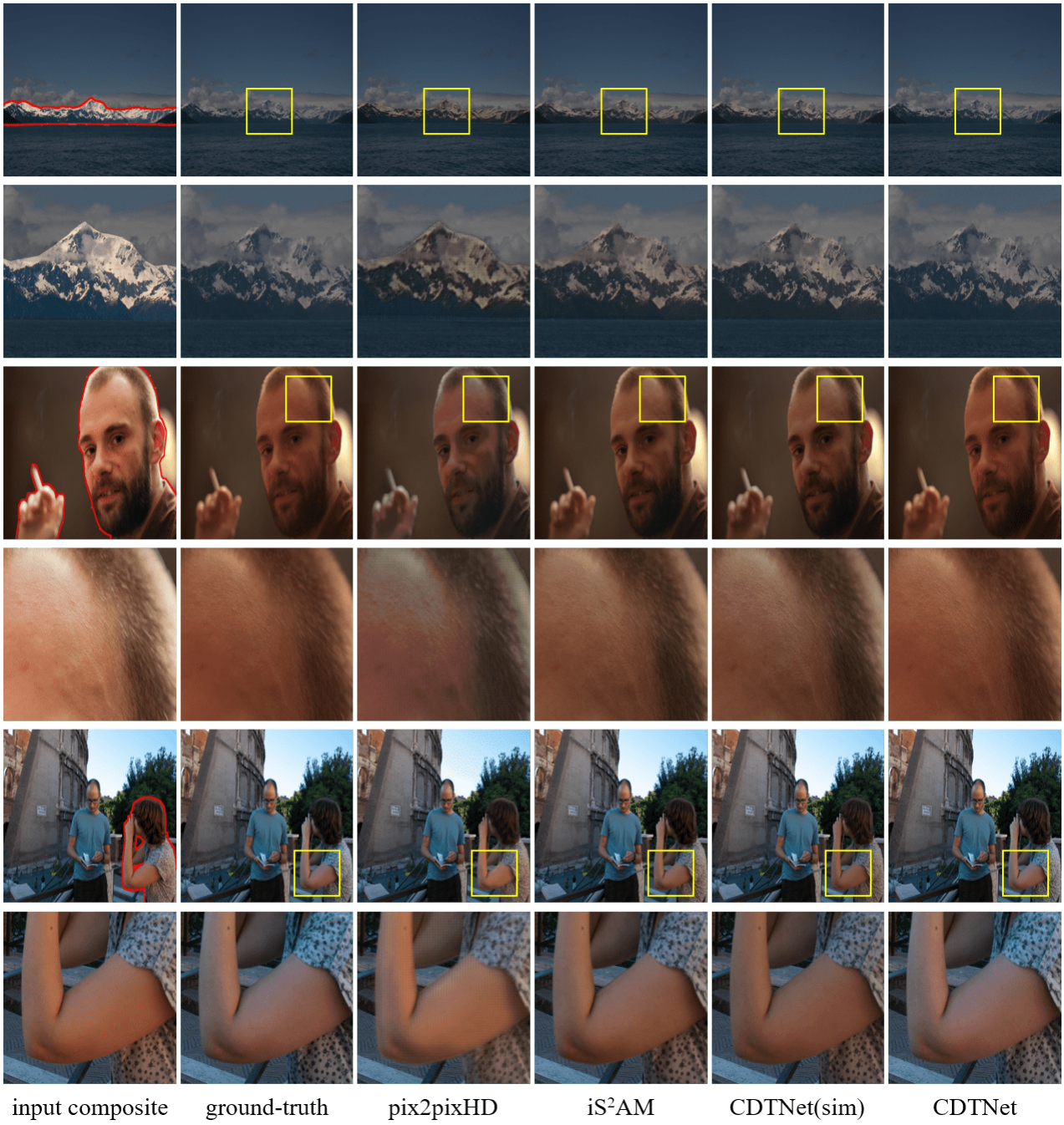

Figure 4. Odd rows show the input composite image, the ground-truth image, as well as example results generated by pix2pixHD [12], iS^{2}AM [10], our simplified variant, CDTNet (sim), which uses deep RGB-to-RGB transformation only, and our CDTNet on 1024×1024 resolution. The red border lines indicate the foreground, and the yellow boxes zoom in the particular regions for a better observation.

8 and Figure 9. Since pix2pixHD is not well-designed for image harmonization, it tends to produce checkerboard artifacts and halo artifacts in the foreground region, which is especially obvious when zooming in. Therefore, the results of pix2pixHD are far from satisfactory in the real scenario. Compared with iS^{2}AM, our CDTNet is more capable of generating harmonious outputs in real scenarios, which demonstrates the superiority and generalizability of our proposed method.

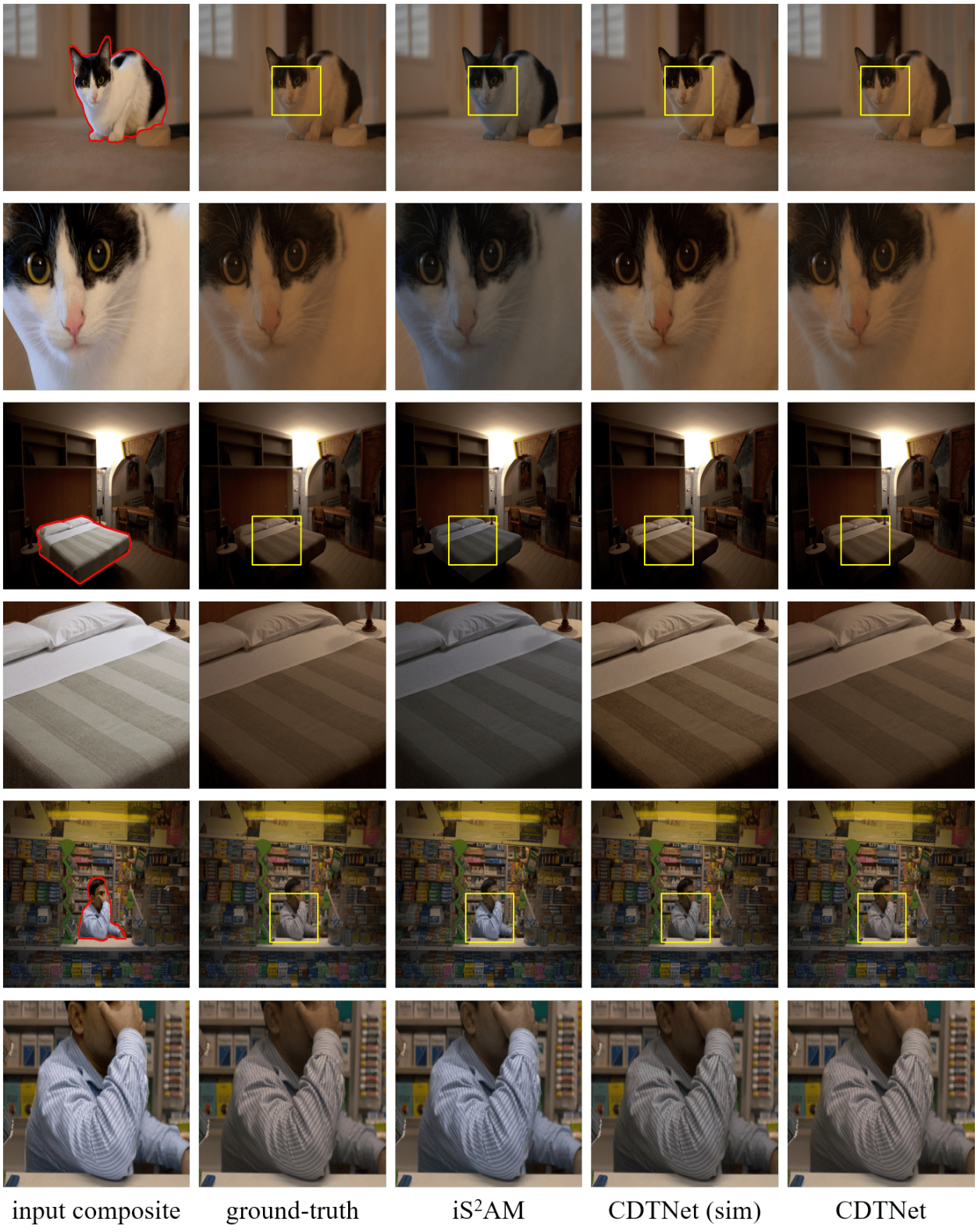

Figure 5. Odd rows show the input composite image, the ground-truth image, as well as example results generated by iS^{2}AM [10], our simplified variant, CDTNet (sim), which uses deep RGB-to-RGB transformation only, and our CDTNet on 2048×2048 resolution. The red border lines indicate the foreground, and the yellow boxes zoom in the particular regions for a better observation.

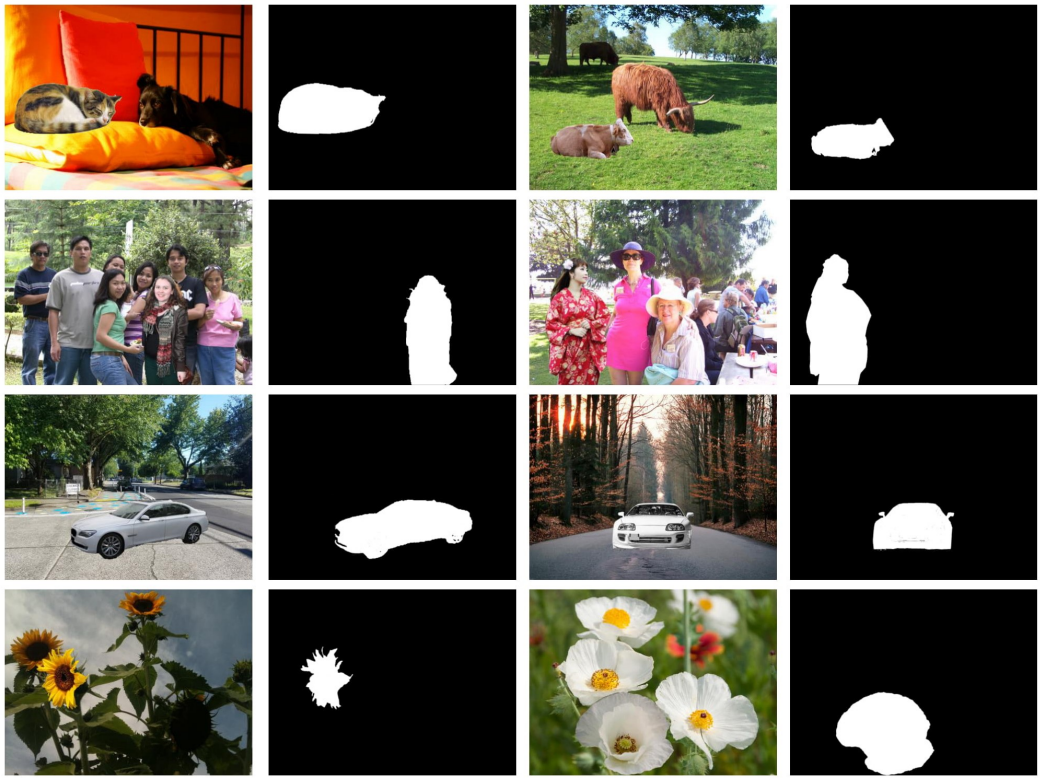

Figure 6. Example composite images and corresponding masks from our created 100 high-resolution real composite images.

Figure 7. A Sample of failure cases on 1024×1024 resolution. From left to right, we show the input composite image, the ground-truth image, as well as the harmonized result generated using only the low-resolution generator, only the color mapping module, and our CDTNet.

7. Limitations

Although our model could achieve stable and effective image harmonization performance on different resolutions, it might encounter failures with unrealistic local results. For example, in Figure 7, the petal in the input composite image is overexposed, which still remain very bright after applying global RGB-to-RGB transformation (see the RGB-to-RGB result $\hat{\mathbf{I}}^{hr}_{rgb}$). The RGB-to-RGB result, as partial input of the refinement module, may adversely affect the refinement module, leading to inharmonious local results.

References

[1] Ivan Anokhin, Pavel Solovev, Denis Korzhenkov, Alexey Kharlamov, Taras Khakhulin, Aleksei Silvestrov, Sergey Nikolenko, Victor Lempitsky, and Gleb Sterkin. High-resolution daytime translation without domain labels. In *CVPR*, 2020. 1

[2] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 1952. 4

[3] Qifeng Chen and Vladlen Koltun. Photographic Image Synthesis with Cascaded Refinement Networks. In *ICCV*, 2017. 1

[4] 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, 4

[5] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Trans. Image Process.*, 29:4759–4771, 2020. 1, 4

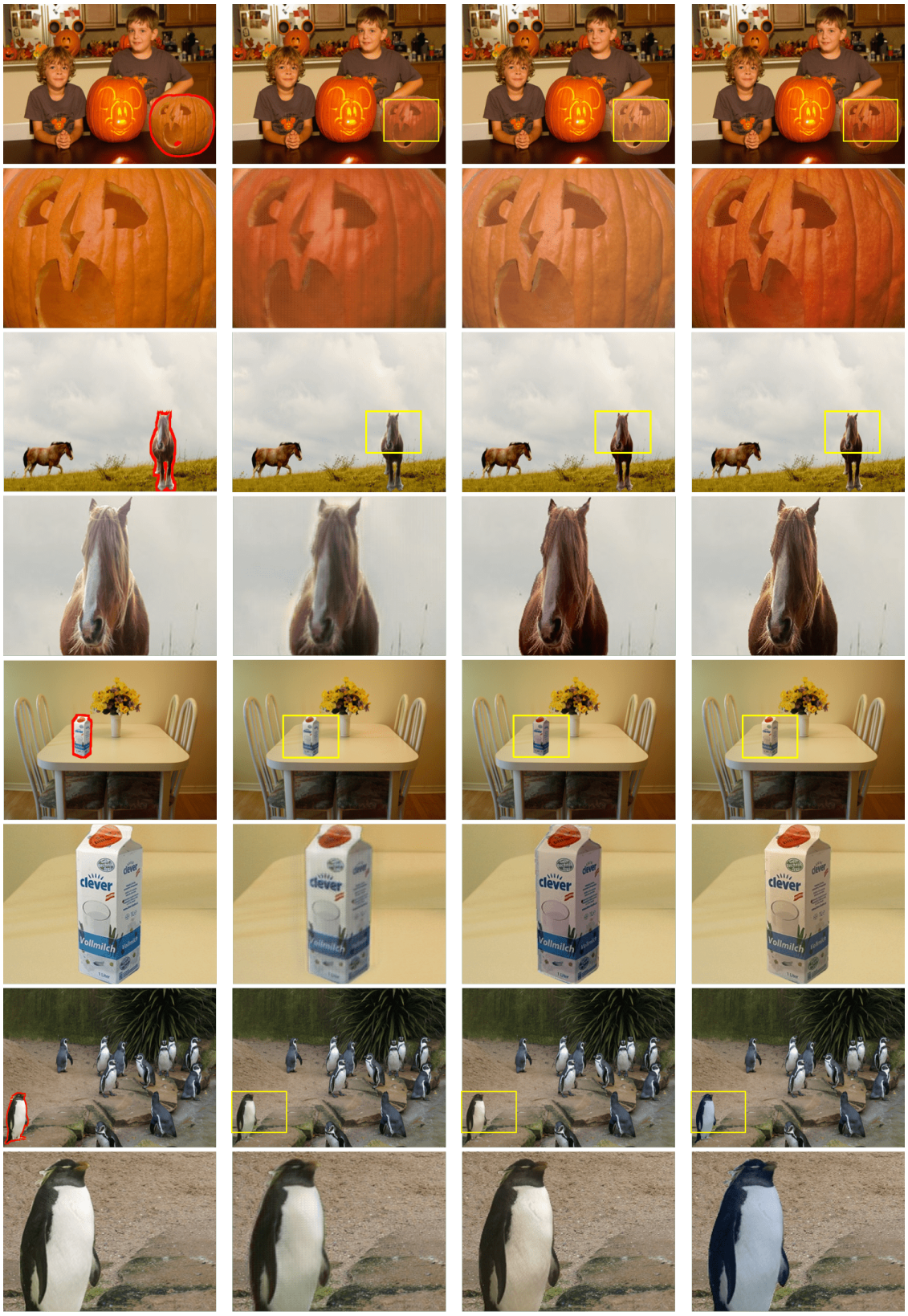

Figure 8. Example results on high-resolution real composite images. From left to right, we show the input composite image and the harmonized results generated by pix2pixHD [12], iS^{2}AM [10], and our CDTNet on 1024×1024 resolution. The red border lines indicate the foreground, and the yellow boxes zoom in the particular regions for a better observation. All images are resized in this figure for better visualization.

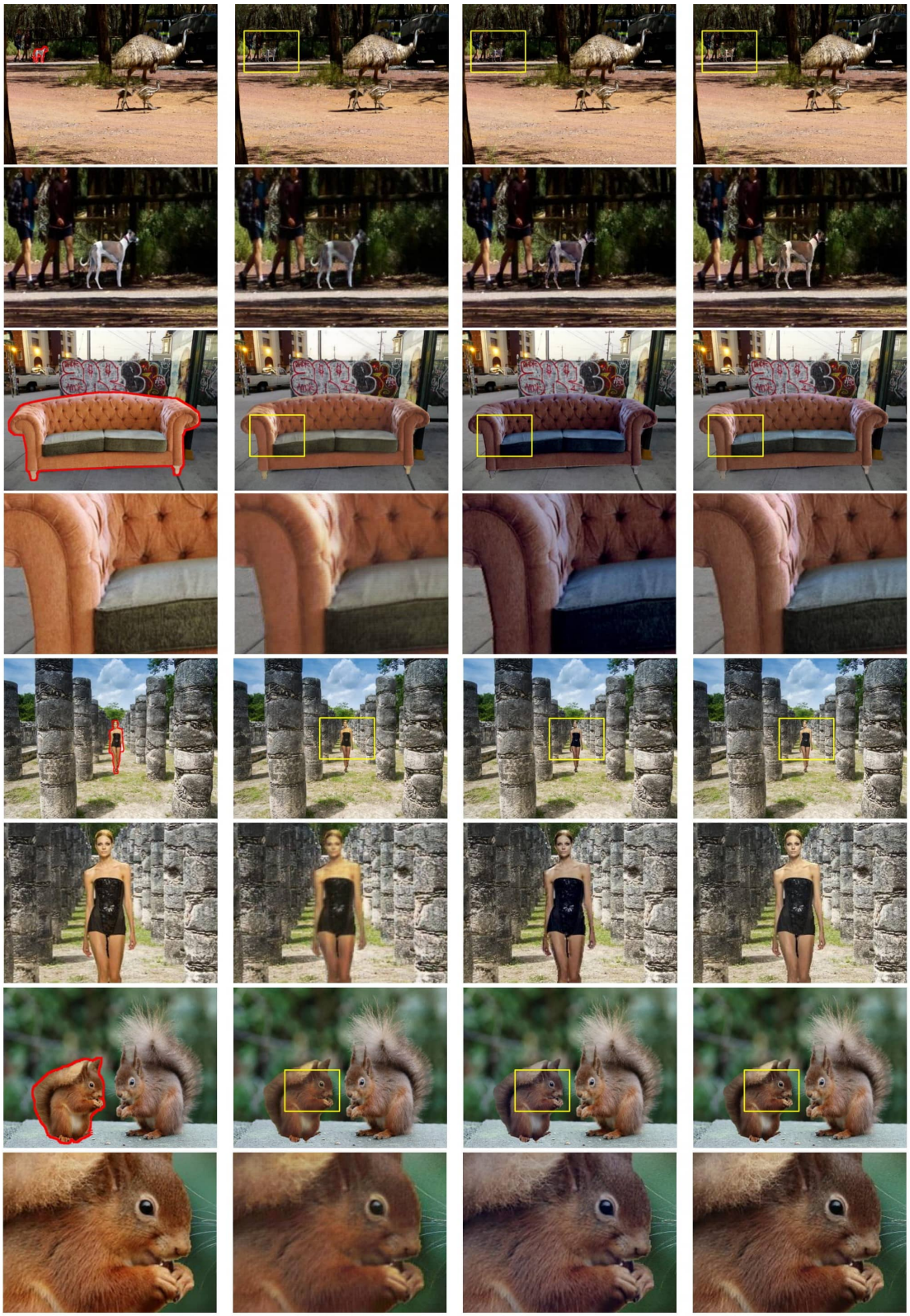

Figure 9. Example results on high-resolution real composite images. From left to right, we show the input composite image and the harmonized results generated by pix2pixHD [12], iS^{2}AM [10], and our CDTNet on 1024×1024 resolution. The red border lines indicate the foreground, and the yellow boxes zoom in the particular regions for a better observation. All images are resized in this figure for better visualization.

[6] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *CVPR*, 2021. 1

[7] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. 4

[8] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *CVPR*, 2016. 4

[9] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *CVPR*, 2021. 1

[10] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *WACV*, 2021. 1, 2, 3, 4, 6, 7, 9, 10

[11] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, 2017. 4

[12] Ting Chun Wang, Ming Yu Liu, Jun Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *CVPR*, 2018. 1, 3, 4, 6, 9, 10

[13] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, June 2018. 1