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

Abstract

Given a composite image, image harmonization aims to adjust the foreground to make it compatible with the background. High-resolution image harmonization is in high demand, but still remains unexplored. Conventional image harmonization methods learn global RGB-to-RGB transformation which could effortlessly scale to high resolution, but ignore diverse local context. Recent deep learning methods learn the dense pixel-to-pixel transformation which could generate harmonious outputs, but are highly constrained in low resolution. In this work, we propose a high-resolution image harmonization network with Collaborative Dual Transformation (CDTNet) to combine pixel-to-pixel transformation and RGB-to-RGB transformation coherently in an end-to-end network. 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. Extensive experiments on high-resolution benchmark dataset and our created high-resolution real composite images demonstrate that our CDTNet strikes a good balance between efficiency and effectiveness. Our used datasets can be found in https://github.com/bcmi/CDTNet-High-Resolution-Image-Harmonization.

1. Introduction

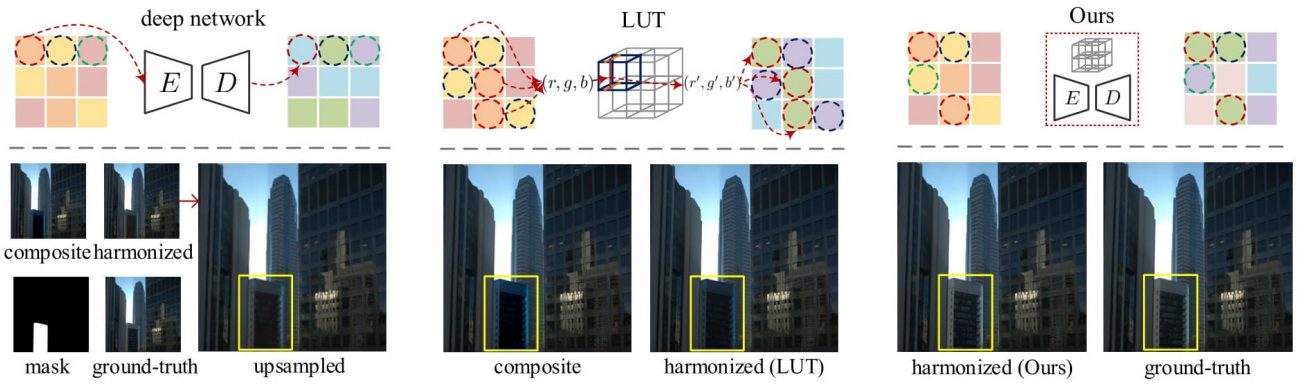
Image composition [26] combines foreground and background from different images into a composite image. The quality of composite image may be degraded by the appearance (*e.g.*, tone, illumination) inconsistency between foreground and background. To address this issue, image harmonization adjusts the foreground appearance to make it compatible with the background. Deep image harmonization methods [7, 8, 15, 24, 31, 34] have achieved remarkable progress by learning the dense pixel-to-pixel transformation between composite images and ground-truth harmonized images. However, they only performed low-resolution (*e.g.*, 256×256) image harmonization, and a naïve upsampling can merely lead to a large yet blurry output (see Figure 1a).

Though directly training with high-resolution images could seemingly address this issue, the computational cost is very expensive. For example, it will cost more than 950G FLOPs (floating point operations) and more than 20 GB memory for [31] to harmonize a 2048×2048 composite image. Besides, the high-resolution network may be weak in capturing long-range dependencies due to local convolution operations [36] (see Section 4.3 and the Supplementary).

Prior to deep image harmonization, conventional harmonization methods [21, 29, 32, 38, 42] mainly used hand-crafted statistical features (*e.g.*, illumination, color temperature, contrast, saturation) to determine color-to-color transformation for foreground adjustment. Color-to-color transformation could be achieved in different color spaces, and we narrow the scope to RGB-to-RGB transformation in this work. RGB-to-RGB transformation is a global transformation of color values. Therefore, it is barely constrained by the number of pixels and could effortlessly scale to high resolution. However, global transformation disregards the local context for each pixel, prone to generate harmonization results with local inharmony.

Our key insight for high-resolution image harmonization is to combine the advantages of both pixel-to-pixel transformation and RGB-to-RGB transformation. We name our high-resolution image harmonization network with Collaborative Dual Transformations (CDT) as CDTNet. 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 combine the best of two worlds. The low-resolution generator is a U-Net [30] structure which takes in a low-resolution composite image and outputs a low-resolution harmonized result. Meanwhile, the encoder feature is used to learn RGB-to-RGB transformation, unlike previous hand-crafted methods [29, 32]. Specifically, we learn several basis lookup tables (LUTs) [10–12, 37] shared by all images and a weight predictor based on the encoder feature to predict image-specific

*Corresponding author.



(a) pixel-to-pixel transformation. (b) RGB-to-RGB transformation. (c) collaborative dual transformations.

Figure 1. In (a), deep harmonization network [31] learns dense transformation for each individual pixel and outputs low-resolution harmonious results, which would be blurry if upsampled. In (b), a 3D lookup table (LUT) in our method learns global color transformation and outputs high-resolution results without considering local context, which might lead to inharmonious local regions. In (c), our full method combines two transformations to achieve the most plausible results. Best viewed by zooming in.

combination coefficients. RGB-to-RGB transformation is performed on high-resolution composite images using the combined LUT. With the RGB-to-RGB result as guidance, the refinement module utilizes both the harmonized result and decoder feature map from the low-resolution generator to compensate for fine-grained local information.

Considering the efficiency of CDTNet, pixel-to-pixel transformation only operates on low-resolution inputs and the refinement module is light-weighted, so the memory cost and computational cost are well-suppressed. Considering the effectiveness of CDTNet, RGB-to-RGB transformation could provide a holistic understanding of the whole image and the sharp edges of the transformed image, while pixel-to-pixel transformation could provide fine-grained local information. As shown in Figure 1b, RGB-to-RGB transformation can obtain a globally reasonable tone and illumination. However, it may produce inharmonious local regions (*i.e.*, the left and right borders of the foreground building) without considering local context. In contrast, our CDTNet can yield visually plausible and harmonious results of high resolution (see Figure 1c). Our contributions could be summarized as follows:

- To the best of our knowledge, this is the first work focusing on high-resolution image harmonization.
- We are the first to achieve deep learning based color-to-color transformation for image harmonization.
- We unify pixel-to-pixel transformation and color-to-color transformation coherently in an end-to-end network named CDTNet.
- Extensive experiments demonstrate that our CDTNet achieves state-of-the-art results with less resource consumption.

2. Related Works

2.1. Image Harmonization

Traditional image harmonization methods [6, 18, 21, 27–29, 32, 33, 38, 42] mainly leveraged color-to-color transformation to match the visual appearance, which could be further divided into non-linear transformations [32, 38] and linear transformations [21, 29, 42]. [32, 38] proposed to match the pyramid histograms or histogram zones to address the appearance inconsistency. [21, 29, 42] applied a simple color adjustment to modify the foreground distribution by shifting and scaling the color values.

Recently, deep image harmonization methods [8, 9, 14, 19, 24] focused on learning dense pixel-to-pixel transformation from deep learning networks. [31, 34] both leveraged auxiliary semantic features to improve the basic image harmonization network. [8] introduced a domain verification discriminator pulling close the foreground domain and background domain. [9, 17] explored various attention mechanisms for image harmonization. [7, 24] explicitly used background domain/style to guide the foreground harmonization. [15] harmonized composite images by harmonizing reflectance and illumination separately. Different from existing methods, we focus on high-resolution image harmonization. Besides, instead of using only one type of transformation, we combine the complementary RGB-to-RGB transformation and pixel-to-pixel transformation into an end-to-end architecture.

2.2. High-Resolution Image-to-Image Translation

To the best of our knowledge, there are no previous works focusing on high-resolution image harmonization, but high-resolution image-to-image translation has been studied in many other fields like image segmentation [5, 22],

image inpainting [39], image matting [40], style transfer [2, 23], and image synthesis [4, 25, 35]. Recent works could be mainly divided into three groups. The first group is placing a low-resolution generator embedded in a high-resolution generator. To name a few, [35] pioneered the embedded scheme and extended the pix2pix to pix2pixHD to adapt to high-resolution applications. A slew of works followed this line and proposed Progressive GAN [20] and its variants [1, 16]. [4] employed a cascade of refinement modules to scale to high resolution. The second group is to stitch/merge low-resolution outputs. [40] cropped the high-resolution image into patches and processed each patch with cross-patch consistency. [2] shifted and downsampled the high-resolution image into multiple low-resolution images for separate processing. The third group leveraged deep learning techniques to predict color transformation [13, 41], which is not constrained by the image resolution. Inspired by the third group of methods, this work is a pioneer in applying deep color-to-color transformation to image harmonization. Besides, collaborating with a low-resolution deep image harmonization network (*i.e.*, pixel-to-pixel transformation) and a refinement module, our network can produce better high-resolution harmonization results with limited resources.

3. Our Method

We propose a novel network, CDTNet, to reduce the computational burden and simultaneously maintain the harmonization performance. The pipeline of our CDTNet is shown in Figure 2. Given a high-resolution composite image $\tilde{\mathbf{I}}^{hr} \in \mathbb{R}^{H\times W\times 3}$ and foreground mask $\mathbf{M}^{hr}$, image harmonization aims to obtain the harmonized result $\hat{\mathbf{I}}^{hr} \in \mathbb{R}^{H\times W\times 3}$. We first downsample $(\tilde{\mathbf{I}}^{hr}, \mathbf{M}^{hr})$ to $h \times w$ (*e.g.*, $h = w = 256$) to obtain low-resolution $(\tilde{\mathbf{I}}^{lr}, \mathbf{M}^{lr})$. Our network contains three parts: a low-resolution generator, a color mapping module, and a light-weighted refinement module. The low-resolution image harmonization network is a U-Net-like architecture with encoder E and decoder D, which takes in the low-resolution $(\tilde{\mathbf{I}}^{lr}, \mathbf{M}^{lr})$ and outputs low-resolution result $\hat{\mathbf{I}}^{lr}_{pix}$. The RGB-to-RGB transformation is built upon several basis transformations (*i.e.*, LUTs) $\{\Phi_n\}_{n=1,\dots,N}$ and a weight predictor. The weight predictor takes the bottleneck feature map $\mathbf{F}_{enc}$ extracted from E as input to predict the combination coefficients of basis transformations. After applying combined transformation, we could get a high-resolution output $\hat{\mathbf{I}}^{hr}_{rgb}$. Then, the upsampled $\hat{\mathbf{I}}^{lr}_{pix}$, the high-resolution output $\hat{\mathbf{I}}^{hr}_{rgb}$, the last decoder feature map $\mathbf{F}_{dec}$ from low-resolution generator, together with foreground mask $\mathbf{M}^{hr}$ are passed through refinement module R to obtain a better harmonization output $\hat{\mathbf{I}}^{hr}$, which is expected to be close to the high-resolution ground-truth image $\mathbf{I}^{hr}$.

3.1. Pixel-to-Pixel Transformation

Dense pixel-to-pixel transformation has been widely explored by recent deep image harmonization methods. Pixel-to-pixel transformation is adept at adjusting each foreground pixel according to its local context, which could hardly be achieved by RGB-to-RGB transformation. To reduce the computational complexity and memory burden, we propose to leverage the deep image harmonization network with downsampled resolution.

The low-resolution harmonization network could be realized by any generator with an encoder-decoder structure. Recently in [31], they propose to apply an image blending layer to the last decoder feature map $\mathbf{F}_{dec}$ to obtain the harmonized foreground and the soft attention mask. Then the final output of the encoder-decoder network is obtained by blending harmonized foreground and input composite background using the soft attention mask. They also propose a blending layer equipped generator named iS^{2}AM. Considering the simple structure and competitive performance of iS^{2}AM, we adopt it as the low-resolution generator responsible for our pixel-to-pixel transformation. As shown in Figure 2, the low-resolution composite image and mask $(\tilde{\mathbf{I}}^{lr}, \mathbf{M}^{lr})$ are concatenated channel-wisely. After passing the input through the generator (with encoder E and decoder D), we enforce the harmonized output $\hat{\mathbf{I}}^{lr}_{pix} = D(E(\tilde{\mathbf{I}}^{lr}, \mathbf{M}^{lr}))$ to be close to the downsampled ground-truth real image $\mathbf{I}^{lr} \in \mathbb{R}^{h\times w\times 3}$ by minimizing the reconstruction loss $\mathcal{L}_{pix} = \|\hat{\mathbf{I}}^{lr}_{pix} - \mathbf{I}^{lr}\|_1$.

3.2. RGB-to-RGB Transformation

The simple hand-crafted features used in traditional image harmonization methods have been proven insufficient to acquire appealing harmonization results [8]. To enable expressive and flexible RGB-to-RGB transformation, we jointly learn a few basis non-linear transformations and a weight predictor to predict the combination coefficient for each transformation.

For basis transformation, we employ lookup tables (LUTs) due to their simplicity and expressiveness. LUT replaces expensive input/output operation with a simple array indexing operation, and it has been employed to transform input color value to desired output color value in many image processing methods [10–12, 37, 41]. As we expect joint control on the RGB values (3 channels as a whole) instead of a single channel, our used LUT could be regarded as a 3-dimensional grid that defines a conversion matrix in RGB space. As shown in Figure 1b, the RGB-to-RGB transformation using LUT could be achieved by two steps. First, given input RGB values (r, g, b), we *look up* its 3D coordinates in LUT. Then *trilinear interpolation* is performed based on its eight nearest surrounding elements to calculate the output RGB values (r', g', b') without sacrificing

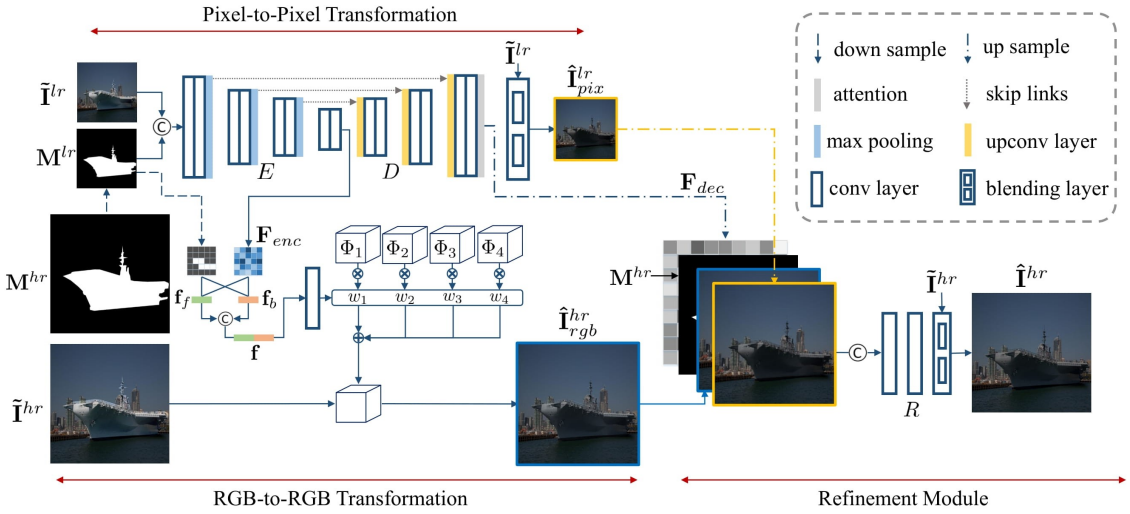


Figure 2. The illustration of our CDTNet for high-resolution image harmonization, which takes in $(\tilde{\mathbf{I}}^{hr}, \mathbf{M}^{hr})$ and outputs $\hat{\mathbf{I}}^{hr}$. CDTNet contains a low-resolution generator (encoder E and decoder D) for pixel-to-pixel transformation, a color mapping module (basis LUTs and weight predictor) for RGB-to-RGB transformation, and a refinement module R.

the continuity of the RGB values. Since the interpolation only introduces a small amount of computation, the combination of these two steps is still computationally efficient.

Different LUTs could have different output colors, so we employ a set of N learnable LUTs $\{\Phi_n\}_{n=1,...,N}$ as the basis transformations to cover the color transformation space between inconsistent foreground and background regions. Once learned, the basis LUTs are universal for all images. Then, inspired by [41], we employ a weight predictor responsible for image-specific transformation by predicting image-specific combination coefficients. Our weight predictor is built upon the bottleneck feature map in low-resolution harmonization network for the following reasons. Firstly, deep network has a strong capability of capturing the context of input images. Thus the extracted bottleneck feature map contains rich information that is useful for image harmonization. Secondly, by sharing the encoder, pixel-to-pixel transformation and RGB-to-RGB transformation are accommodated under a multi-task learning framework, in which two tasks can benefit each other.

As shown in Figure 2, the low-resolution generator takes in low-resolution image and mask, and extracts the bottleneck feature map $\mathbf{F}_{enc} = E(\tilde{\mathbf{I}}^{lr}, \mathbf{M}^{lr})$. Since the predicted coefficients are expected to adjust the foreground according to the background, we conjecture that explicitly comparing foreground and background features may help learn better coefficients (see Table 4). In particular, based on the mask $\mathbf{M}^{lr}$, we use average pooling to aggregate the L-dimensional feature vector $\mathbf{f}_f$ and $\mathbf{f}_b$ for foreground and background separately, which are concatenated as a $2L$-dimensional feature vector $\mathbf{f}$. Then, we apply a simple fully connected layer to $\mathbf{f}$ to obtain the image-specific coefficients $\{w_n\}_{n=1,...,N}$, where N is the number of basis LUTs. For a high-resolution composite image $\tilde{\mathbf{I}}^{hr}$, given N learnable LUTs $\{\Phi_n\}_{n=1,...,N}$ and combination coefficients, its harmonization output of RGB-to-RGB transformation is obtained as

$$\hat{\mathbf{I}}^{hr}_{rgb} = (\sum_{n=1}^{N} w_n \Phi_n)(\tilde{\mathbf{I}}^{hr}). \tag{1}$$

Note that in RGB-to-RGB transformation, the background also remains the same after harmonization. To enforce the high-resolution output to be close to the high-resolution ground-truth real image $\mathbf{I}^{hr}$, we employ the reconstruction loss $\mathcal{L}_{rgb} = \|\hat{\mathbf{I}}^{hr}_{rgb} - \mathbf{I}^{hr}\|_1$.

3.3. Light-weighted Refinement

After RGB-to-RGB transformation and pixel-to-pixel transformation, we can obtain RGB-to-RGB result $\hat{\mathbf{I}}^{hr}_{rgb} \in \mathbb{R}^{H\times W\times 3}$ and pixel-to-pixel result $\hat{\mathbf{I}}^{lr}_{pix} \in \mathbb{R}^{h\times w\times 3}$. $\hat{\mathbf{I}}^{hr}_{rgb}$ is with high-resolution but insufficient local context. $\hat{\mathbf{I}}^{lr}_{pix}$ is with low-resolution but rich local context. Hence, they are complementary with each other. However, naively mixing $\hat{\mathbf{I}}^{hr}_{rgb}$ and $\hat{\mathbf{I}}^{lr}_{pix}$ can only lead to a blurry and unsatisfactory output, so we design a light-weighted refinement module R to generate better high-resolution result $\hat{\mathbf{I}}^{hr} \in \mathbb{R}^{H\times W\times 3}$.

Specifically, we first concatenate bilinearly upsampled $Up(\hat{\mathbf{I}}^{lr}_{pix}) \in \mathbb{R}^{H\times W\times 3}$ and $\hat{\mathbf{I}}^{hr}_{rgb} \in \mathbb{R}^{H\times W\times 3}$ in channels. As stated in [7], mask is essential for image harmonization by explicitly indicating the foreground region. Additionally, the last decoder feature map $\mathbf{F}_{dec}$ accounts for both harmonized foreground and soft attention mask, so we conjecture

that $\mathbf{F}_{dec}$ contains rich prior knowledge of both harmonization and local context. Therefore, we further append the binary mask $\mathbf{M}^{hr} \in \mathbb{R}^{H\times W\times 1}$ and bilinearly upsampled $Up(\mathbf{F}_{dec}) \in \mathbb{R}^{H\times W\times c}$ to the input.

Our refinement module R contains two convolution layers with kernel 3 and stride 1, each followed by a batch normalization and an ELU. At the end of R, we also employ an image blending layer to blend both the high-resolution harmonized foreground and input composite image $\tilde{\mathbf{I}}^{hr}$. Consequently, the refinement module R receives the $H \times W \times (c+7)$ input and produces the better refined high-resolution output $\hat{\mathbf{I}}^{hr} \in \mathbb{R}^{H\times W\times 3}$. We impose the reconstruction loss to enforce the $\hat{\mathbf{I}}^{hr}$ to be close to the high-resolution ground-truth real image $\mathbf{I}^{hr}$, which is denoted by $\mathcal{L}_{ref} = \|\hat{\mathbf{I}}^{hr} - \mathbf{I}^{hr}\|_1$.

Therefore, the overall loss function of our CDTNet is

$$\mathcal{L} = \mathcal{L}_{pix} + \mathcal{L}_{rgb} + \mathcal{L}_{ref}. \tag{2}$$

Despite the simple structure of the refinement module, it performs well in fusing the RGB-to-RGB result and the pixel-to-pixel result due to the informative input and the blending layer, which will be validated in Table 4.

4. Experiments

4.1. Dataset Statistics

Existing image harmonization benchmark iHarmony4 [8] is not specially constructed for high-resolution image harmonization. Among four sub-datasets, HCOCO, HFlickr, and Hday2night are all at a resolution lower than 1024, and only HAdobe5k contains high-resolution images. Therefore, we conduct experiments on HAdobe5k, which contains 19437 training and 2160 testing pairs of high-resolution composite images and real images.

Considering that the composite images in HAdobe5k are synthesized composite images, we follow [8, 34] and further evaluate our model on real composite images. However, existing 99 real composite images [34, 38] are also at a low resolution. Thus, we collect images from Open Image Dataset V6 [3] and Flickr [1], and create 100 high-resolution real composite images with diverse foregrounds and backgrounds (see the Supplementary).

4.2. Implementation Details

As mentioned in Section 3.1, we adopt iS^{2}AM backbone proposed in [31] as the low-resolution generator. In the color mapping module, $\mathbf{f}$ is a 512-dimensional feature vector with $L = 256$. The basis transformations contain $N = 4$ learnable LUTs. We also investigate the impact of using different N in Supplementary. To ensure that color transformation is well-behaved, we clip the transformed RGB value into the range $[0, 1]$. In the refinement module, the number of input channels is 39 with $c = 32$. Our network is implemented using Pytorch 1.6.0 and trained using Adam optimizer with learning rate of $1e{-}4$ on ubuntu 18.04 LTS operation system, with 64GB memory, Intel Core i7-8700K CPU, and two GeForce GTX 1080 Ti GPUs. **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) by default.**

We employ four metrics for harmonization performance evaluation, including MSE, foreground MSE (fMSE), PSNR, and SSIM, as well as three metrics for efficiency evaluation, including the average inference time per image, memory cost, and FLOPs. The average inference time per image is evaluated on a single NVIDIA GTX 1080 Ti GPU, and the memory cost and FLOPs are estimated with the network analyzer "torchstat".

4.3. Comparison with Existing Methods

Since there are no existing methods specifically designed for high-resolution image harmonization, we compare two baseline groups. The first group contains five recent low-resolution image harmonization methods [8, 9, 15, 24, 31]. The second group contains high-resolution image-to-image translation methods. We select three representative methods pix2pixHD [35], HiDT [2], and CRN [4] for comparison. For fairness, we transplant all baselines to high-resolution image harmonization with the slightest modification of their officially released code (see the Supplementary). In this section, we refer to our CDTNet with the resolution of low-resolution generator being 256 (*resp.*, 512) as CDTNet-256 (*resp.*, CDTNet-512). Moreover, we build a simplified variant (sim) of our CDTNet, which has the same training procedure but only uses deep RGB-to-RGB transformation during inference.

First, we evaluate the harmonization performance of different methods on 1024×1024 resolution. In Table 1, the performances of HiDT and CRN are poor, probably because merging multiple shifted low-resolution results may disturb the pixel values, and generating from repeated refinement may amplify the artifacts. Since high-resolution image-to-image translation baselines are not well-designed for image harmonization task, their overall performance is far from satisfactory. Among the image harmonization baselines, iS^{2}AM achieves competitive performance, which coincides with its superiority at low resolution as reported in [31] and reveals the reason for employing iS^{2}AM as the low-resolution generator in our method. However, the high-resolution network for pixel-to-pixel transformation may not be good at capturing long-range dependency due to local convolution operations [36], especially for the images with

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

Image Size	Method	MSE↓	PSNR↑	fMSE↓	SSIM↑
1024 × 1024	Composite images	352.05	28.10	2122.37	0.9642
	pix2pixHD [35]	63.45	31.64	332.43	0.9135
	CRN [4]	90.11	29.77	259.28	0.8225
	HiDT [2]	265.32	29.95	1501.93	0.9628
	DoveNet [8]	51.00	34.81	312.88	0.9729
	S^2AM [9]	47.01	35.68	262.39	0.9784
	Guo et al. [15]	56.34	34.69	417.33	0.9471
	RainNet [24]	42.56	36.61	305.17	0.9844
	iS^2AM [31]	25.03	38.29	168.85	0.9846
	CDTNet-256 (sim)	31.15	37.65	195.93	0.9841
	CDTNet-256	**21.24**	**38.77**	**152.13**	**0.9868**

Image Size	Method	MSE↓	PSNR↑	fMSE↓	SSIM↑
2048 × 2048	Composite images	353.92	28.07	2139.97	0.9631
	iS^2AM [31]	46.37	36.57	271.59	0.9838
	CDTNet-256 (sim)	41.11	37.28	234.06	0.9819
	CDTNet-256	29.02	37.66	198.85	0.9845
	CDTNet-512 (sim)	38.31	37.05	233.44	0.9828
	CDTNet-512	**23.35**	**38.45**	**159.13**	**0.9853**

Table 1. Quantitative harmonization performance evaluation of different methods. "CDTNet-256/512" means the resolution of the low-resolution generator is $256 \times 256/512 \times 512$, and "(sim)" represents the simplified variant with deep RGB-to-RGB transformation only.

Image Size	Method	HCOCO		HAdobe5k		HFlickr		Hday2night		All	
		MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑
256×256	iS^2AM [31]	16.48	39.16	22.60	37.24	69.67	33.56	40.59	37.72	24.65	37.95
	CDTNet-256	16.25	39.15	20.62	38.24	68.61	33.55	36.72	37.95	23.75	38.23

Table 2. Quantitative comparison on low-resolution (256×256) image harmonization on iHarmony4 dataset. The resolution of our low-resolution generator is 256×256. The performance of iS^2AM is tested using the publicly released model from [31].

large foregrounds (*e.g.*, row 1 in Figure 3). The detailed results on different foreground ratio ranges can be found in the Supplementary. Our CDTNet-256 outperforms [8, 9, 15, 24] by a large margin and also beats iS^2AM. Even our simplified variant CDTNet-256(sim) outperforms most methods, which demonstrates the expressiveness of our proposed deep RGB-to-RGB transformation.

With the observation that iS^2AM [31] is the most competitive baseline, we further compare with iS^2AM on 2048×2048 resolution in Table 1. The advantage of our method is more obvious, and CDTNet-512 significantly outperforms iS^2AM. When the resolution of the low-resolution generator is reduced to 256, our CDTNet-256 still achieves competitive performance and exceeds iS^2AM by a large margin on 2048×2048 resolution. For the simplified variants, CDTNet-256 (sim) and CDTNet-512 (sim) both exceed iS^2AM, making our proposed deep RGB-to-RGB transformation a strong competitor to high-resolution pixel-to-pixel transformation.

Additionally, to evaluate the performance on low-resolution image harmonization, we also compare with the strongest baseline iS^2AM on 256×256 resolution images by using the training set and test set from iHarmony4 dataset [8]. The results in Table 2 show that our CDTNet is slightly better than iS^2AM.

In Table 3, we report the inference time, memory cost, and FLOPs when harmonizing a single image in the testing stage. Compared to iS^2AM, our CDTNet-256 (*resp.*, CDTNet-512) is 26.53% (*resp.*, 63.02%) faster with 62.65% less memory cost and 67.40% fewer FLOPs for 1024×1024 (*resp.*, 2048×2048) images. When the resolution of low-resolution generator is reduced to 256 on 2048×2048, CDTNet-256 requires even less computational resources and costs less time. For our simplified variants, the superiority of efficiency is more striking, saving $> 95\%$ memory and FLOPs and reducing $> 75\%$ time, which also demonstrates that RGB-to-RGB transformation is not constrained by the number of pixels and could achieve higher efficiency and lower memory consumption.

4.4. Ablation Studies

Recall that 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. Therefore, in this section, we demonstrate the role of low-resolution generator, color mapping module, and refinement module by ablating each component and analyzing different variants with different input types. By taking the 1024×1024 resolution as an example, we report the results in Table 4. We can observe that when we only use a low-resolution generator and naively upsample the low-resolution output, it will lead to unsatisfactory and blurry outputs (row 1). When only using color mapping module, the result (row 2) is much better than upsampled low-resolution output (row 1), which demonstrates the effectiveness of deep RGB-to-RGB transformation. Furthermore, we compare with two variants of color mapping module. Recall that our weight predictor is based on the encoder feature map in low-resolution generator. The first variant is using a separate encoder to extract features for the weight predictor. After using a separate encoder, the performance is downgraded by a large mar-

Image Size	Method	Time↓ (ms)	Reduction	Memory↓ (MB)	Reduction	FLOPs↓ (G)	Reduction
1024 × 1024	iS^{2}AM	14.7	-	5148	-	239.36	-
	CDTNet-256 (sim)	3.5	76.20%	144	97.20%	3.49	98.54%
	CDTNet-256	10.8	26.53%	1923	62.65%	78.05	67.40%
2048 × 2048	iS^{2}AM	31.1	-	20592	-	957.43	-
	CDTNet-256 (sim)	3.5	88.74%	288	98.60%	3.49	99.64%
	CDTNet-256	10.9	64.95%	6723	67.35%	267.10	72.10%
	CDTNet-512 (sim)	3.7	88.10%	574	97.21%	13.95	98.54%
	CDTNet-512	11.5	63.02%	7691	62.65%	312.19	67.40%

Table 3. Quantitative efficiency comparison between iS^{2}AM [31] and our CDTNet.

Row	Pixel Trans	RGB Trans	Refinement	PSNR↑	fMSE↓
1	✓			29.41	265.2
2		✓		37.65	195.93
3		w/o shared E		36.86	248.82
4		w/o $\mathbf{f}_f \circ \mathbf{f}_b$		37.31	217.91
5	✓	✓	w/o $\hat{\mathbf{I}}^{hr}_{rgb}$	37.29	226.17
6	✓	✓	w/o $\hat{\mathbf{I}}^{lr}_{pix}$	37.78	188.97
7	✓	✓	w/o $\mathbf{M}^{hr}$	37.33	202.14
8	✓	✓	w/o $\mathbf{F}_{dec}$	37.64	193.90
9	✓	✓	w/o B_r	28.13	688.55
10	✓	✓	✓	**38.77**	**152.13**

Table 4. Ablation studies of low-resolution generator, color mapping module, and refinement module on the 1024 × 1024 resolution. "∘" stands for the concatenation, and B_r stands for the blending layer in the refinement module.

gin (row 3 *v.s.* row 2), demonstrating the effectiveness of jointly performing low-resolution image harmonization and learning RGB-to-RGB transformation. The second variant is using globally pooled encoder feature instead of concatenating the pooled foreground and background features. The performance also drops (row 4 *v.s.* row 2), which shows the advantage of dealing with foreground and background separately.

Then, we ablate each type of input for the refinement module. We ablate the RGB-to-RGB result $\hat{\mathbf{I}}^{hr}_{rgb}$, the upsampled pixel-to-pixel result $\hat{\mathbf{I}}^{lr}_{pix}$, mask $\mathbf{M}^{hr}$, and upsampled low-resolution feature $\mathbf{F}_{dec}$, separately (row 5 to row 8). We can see that the results after removing each type of input all become worse, which indicates the necessity of using all types of input (row 10). Besides, we also remove the blending layer in the refinement module. The obtained result (row 9) is significantly downgraded, which shows that blending layer is very useful for maintaining the background and focusing on adjusting the foreground.

To better demonstrate the effectiveness of each component, we provide some example images harmonized by different ablated versions in the Supplementary. Moreover, we investigate the efficiency of each individual module and report the quantitative results (time, memory, FLOPs as in Table 3) in the Supplementary.

4.5. Qualitative Analyses

We show the high-resolution (1024 × 1024) harmonization results of pix2pixHD [35], iS^{2}AM [31], and our CDTNet in Figure 3. pix2pixHD is not specifically designed for image harmonization, so its performance is less satisfactory. Due to the weak ability to capture long-range dependencies, iS^{2}AM may fail to generate globally harmonious foreground. Compared with them, the results of our model are more plausible and harmonious, which are visually closer to the ground-truth image. More results and analyses are left to the Supplementary.

4.6. Evaluation on Real Composite Images

As mentioned in Section 4.1, to evaluate the effectiveness of our proposed CDTNet in real scenarios, we manually create 100 high-resolution real composite images and conduct user study to compare our CDTNet with pix2pixHD [35] and iS^{2}AM [31] following [8, 34]. The details of user study and harmonization results on real composite images could be found in the Supplementary.

5. Conclusion

In this work, we have proposed a novel high-resolution image harmonization method CDTNet with collaborative dual transformations. Our CDTNet consists of a low-resolution generator, a color mapping module, and a refinement module, which integrates pixel-to-pixel transformation and RGB-to-RGB transformation into a unified end-to-end network. Extensive experiments on HAdobe5k dataset and real composite images have demonstrated that our method can achieve better harmonization performance with higher efficiency.

Acknowledgement

The work is supported by the National Key R&D Program of China (2018AAA0100704), the National Science Foundation of China (62076162), and the Shanghai Municipal Science and Technology Major Project, China (2021SHZDZX0102).

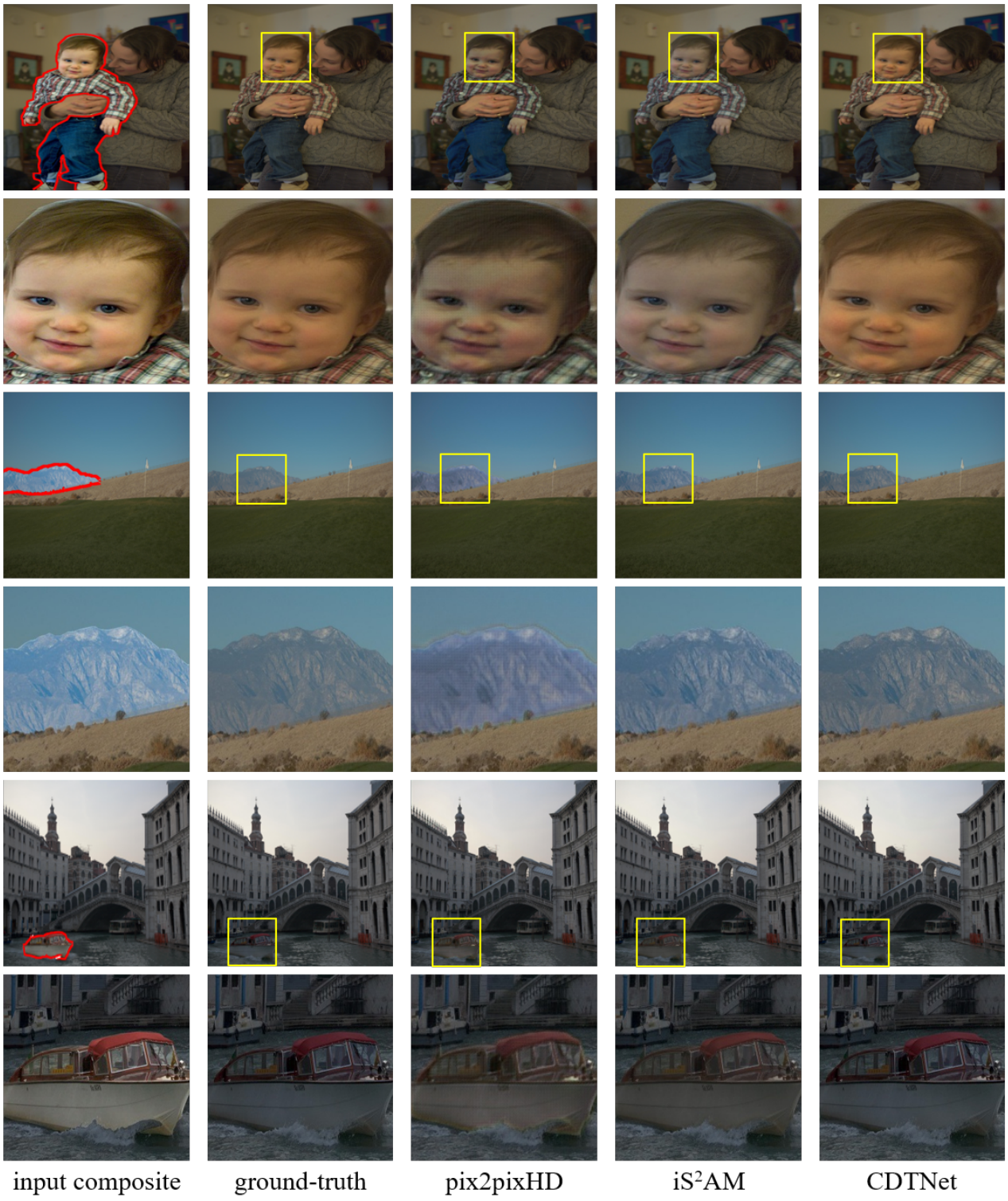


Figure 3. The odd rows show the input composite image, the ground-truth image, as well as example results generated by pix2pixHD [35], iS^{2}AM [31], and our CDTNet on 1024×1024 resolution. The red border lines indicate the foreground, and the yellow boxes zoom in the particular regions for a better observation.

References

[1] Paolo Andreini, Simone Bonechi, Monica Bianchini, Alessandro Mecocci, Franco Scarselli, and Andrea Sodi. A Two Stage GAN for High Resolution Retinal Image Generation and Segmentation. *arXiv preprint arXiv:1907.12296*, 2019. 3

[2] 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. 3, 5, 6

[3] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *CVPR*, 2019. 5

[4] Qifeng Chen and Vladlen Koltun. Photographic Image Synthesis with Cascaded Refinement Networks. In *ICCV*, 2017. 3, 5, 6

[5] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *CVPR*, 2019. 2

[6] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. Color harmonization. *ACM Transactions on Graphics*, 25(3):624–630, 2006. 2

[7] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. BargainNet: Background-guided domain translation for image harmonization. In *ICME*, 2021. 1, 2, 4

[8] 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, 2, 3, 5, 6, 7

[9] 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. 2, 5, 6

[10] M. Elad, B. Matalon, and M. Zibulevsky. Image denoising with shrinkage and redundant representations. In *CVPR*, 2006. 1, 3

[11] Bo Fan, Fugen Zhou, and Hongbin Han. Medical image enhancement based on modified lut-mapping derivative and multi-scale layer contrast modification. In *ICSIP*, 2011. 1, 3

[12] Shu Fujita, Norishige Fukushima, Makoto Kimura, and Yutaka Ishibashi. Randomized redundant dct: Efficient denoising by using random subsampling of dct patches. In *SIGGRAPH*, 2015. 1, 3

[13] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 3

[14] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14870–14879, October 2021. 2

[15] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *CVPR*, 2021. 1, 2, 5, 6

[16] Koichi Hamada, Kentaro Tachibana, Tianqi Li, Hiroto Honda, and Yusuke Uchida. Full-body high-resolution anime generation with progressive structure-conditional generative adversarial networks. In *ECCV Workshop on Computer Vision for Fashion, Art and Design*, 2018. 3

[17] Guoqing Hao, Satoshi Iizuka, and Kazuhiro Fukui. Image harmonization with attention-based deep feature modulation. In *BMVC*, 2020. 2

[18] Jiaya Jia, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Drag-and-drop pasting. *ACM Transactions on Graphics*, 25(3):631–637, 2006. 2

[19] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. In *ICCV*, 2021. 2

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 3

[21] Jean-François Lalonde and Alexei A. Efros. Using color compatibility for assessing image realism. In *ICCV*, 2007. 1, 2

[22] G. Lin, A. Milan, C. Shen, and I. Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 2

[23] Tianwei Lin, Zhuoqi Ma, Fu Li, Dongliang He, Xin Li, Errui Ding, Nannan Wang, Jie Li, and Xinbo Gao. Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. *arXiv preprint arXiv:2104.05376*, 2021. 3

[24] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *CVPR*, 2021. 1, 2, 5, 6

[25] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Self-supervised sketch-to-image synthesis. In *AAAI*, 2021. 3

[26] Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. Making images real again: A comprehensive survey on deep image composition. *arXiv preprint arXiv:2106.14490*, 2021. 1

[27] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Transactions on Graphics*, 22(3):313–318, 2003. 2

[28] Francois Pitie, Anil C Kokaram, and Rozenn Dahyot. N-dimensional probability density function transfer and its application to color transfer. In *ICCV*, 2005. 2

[29] Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001. 1, 2

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1

[31] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *WACV*, 2021. 1, 2, 3, 5, 6, 7, 8

[32] Kalyan Sunkavalli, Micah K. Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM Transactions on Graphics*, 29(4):125:1–125:10, 2010. 1, 2

[33] Michael W. Tao, Micah K. Johnson, and Sylvain Paris. Error-tolerant image compositing. In *ECCV*, 2010. 2

[34] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, 2017. 1, 2, 5, 7

[35] 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. 3, 5, 6, 7, 8

[36] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, June 2018. 1, 5

[37] Zhi-Qiang Wen, Yong-Le Lu, Zhi-Gao Zeng, Wen-Qiu Zhu, and Jun-Hua Ai. Optimizing template for lookup-table inverse halftoning using elitist genetic algorithm. *IEEE Signal Processing Letters*, 22(1):71–75, 2015. 1, 3

[38] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly E. Rushmeier. Understanding and improving the realism of image composites. *ACM Transactions on Graphics*, 31(4):84:1–84:10, 2012. 1, 2, 5

[39] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *CVPR*, 2020. 3

[40] Haichao Yu, Ning Xu, Zilong Huang, Yuqian Zhou, and Humphrey Shi. High-resolution deep image matting. In *AAAI*, 2021. 3

[41] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning Image-adaptive 3D Lookup Tables for High Performance Photo Enhancement in Real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):1–1, 2020. 3, 4

[42] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Learning a discriminative model for the perception of realism in composite images. In *ICCV*, 2015. 1, 2