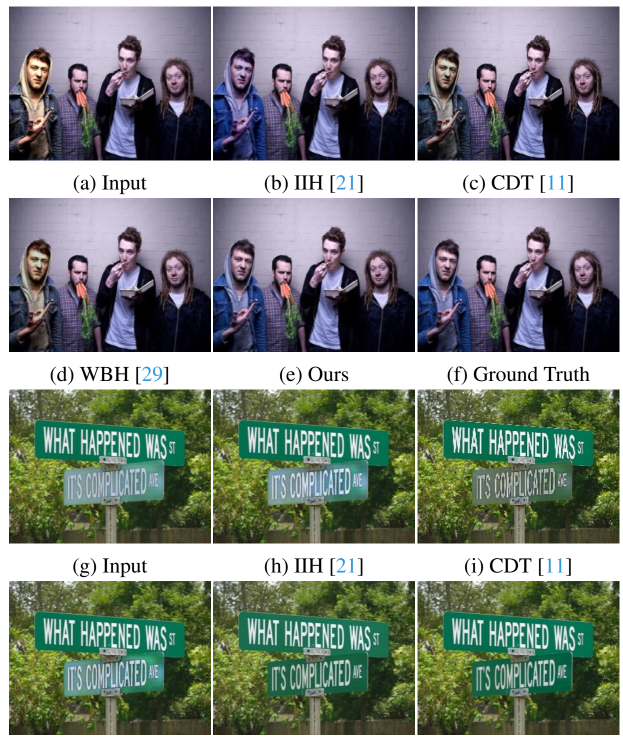
Learning Image Harmonization in the Linear Color Space

Ke Xu, Gerhard Petrus Hancke, Rynson W.H. Lau
City University of Hong Kong
{kkangwing, gp.hancke, rynson.lau}@cityu.edu.hk

Abstract

Harmonizing cut-and-paste images into perceptually realistic ones is challenging, as it requires a full understanding of the discrepancies between the background of the target image and the inserted object. Existing methods mainly adjust the appearances of the inserted object via pixel-level manipulations. They are not effective in correcting color discrepancy caused by different scene illuminations and the image formation processes. We note that image colors are essentially camera ISP projection of the scene radiance. If we can trace the image colors back to the radiance field, we may be able to model the scene illumination and harmonize the discrepancy better. In this paper, we propose a novel neural approach to harmonize the image colors in a camera-independent color space, in which color values are proportional to the scene radiance. To this end, we propose a novel image unprocessing module to estimate an intermediate high dynamic range version of the object to be inserted. We then propose a novel color harmonization module that harmonizes the colors of the inserted object by querying the estimated scene radiance and re-rendering the harmonized object in the output color space. Extensive experiments demonstrate that our method outperforms the state-of-the-art approaches.



(a) Input (b) IIH [21] (c) CDT [11]
(d) WBH [29] (e) Ours (f) Ground Truth
(g) Input (h) IIH [21] (i) CDT [11]
(j) WBH [29] (k) Ours (l) Ground Truth

Figure 1. Harmonization results on the iHarmony4 dataset [12]. Existing harmonization methods tend to produce dull (b-d) or inconsistent (h-j) colors. Our method traces back to and harmonizes the colors in an intermediate linear color space, resulting in more realistic composite images as shown in (e) and (k).

1. Introduction

Image compositing is a common process in vision and graphics. It is a technique to render a novel image by inserting a target object from the source image onto a target image. However, humans can easily identify this cut-and-paste (or composite) image as a synthetic one due to its color [7] and texture inconsistencies [31, 63]. Hence, there is a line of research to develop algorithms to harmonize cut-and-paste images to produce visually realistic output images.

Existing image harmonization methods typically fall into two categories, *i.e.*, non-deep learning-based methods [42, 10, 47, 58, 63] and deep learning-based methods [48, 13, 8, 12, 33, 21, 27, 20]. Non-deep learning based methods try to manipulate low-level image statistics (*e.g.*, textures [47] and colors [42, 63, 58]) of the inserted object, to match with those of the background. These methods often produce unrealistic images of inconsistent colors/textures when the hand-crafted features fail to represent the foreground/background. In contrast, deep learning based methods offer strong capability of modelling region appearances to facilitate harmonization. Some methods explore different priors (*e.g.*, semantics [48], and gradient/color consistency [8, 53]) to constrain the harmonization process. Some other methods [12, 33, 21] may formulate the image harmonization process as a foreground-background transfer learning task.

Despite their success, existing harmonization methods may still produce pale (Figure 1(b-d)) or inconsistent colors (Figure 1(h-j)) across the foreground and background regions, resulting in visually unpleasant images. We note that all these methods model the color harmonization process in the camera output sRGB (*i.e*., low dynamic range) color space. However, object colors in an image are determined not only by their material reflectance and scene illumination, but also by the black-box imaging pipeline (ISP) of the camera. Due to the non-linear operations (*e.g*., tone mapping) within the ISP, pixel intensities of camera output sRGB images are not proportional to the scene radiance, making them unreliable for use in estimating the scene illumination for color harmonization.

To address this problem, we propose in this paper a novel approach to harmonize a cut-and-paste image (captured in low dynamic range) in the high dynamic range domain. Our key idea is to harmonize the scene illumination discrepancies in an intermediate (high dynamic range) color space, in which the scene illumination is proportional to the original scene radiance. To this end, we propose a novel neural network that first converts the source image (containing the target object) into an intermediate high dynamic range domain, then performs the harmonization process, and finally converts the harmonized image back to the low dynamic range sRGB space. To avoid exhaustive modeling of camera-dependent operations, we propose a novel image unprocessing module to estimate a high dynamic range version of the input image in the linear camera-independent CIE XYZ color space. We formulate this image unprocessing process as a diffusion process. We propose a novel color harmonization method that models image colors in the estimated linear color space to produce the final harmonized results. As shown in Figure 1(e,k), our method is able to produce more visually pleasing results. We conduct extensive experiments to demonstrate that our method outperforms state-of-the-art harmonization approaches.

In summary, this paper has three main contributions:

- We propose a novel neural approach for image harmonization that performs the color harmonization process in the linear color space, allowing object color modeling based on faithful scene radiance.
- Our approach includes two novel modules: (1) a novel image unprocessing module to convert the source image (of the target object) into a version in the high dynamic range linear color space, and (2) a novel color harmonization module to harmonize object colors by querying scene radiance information and re-render the harmonized objects in the output color space.
- Extensive experiments show that the proposed method outperforms state-of-the-art harmonization methods.

2. Related Work

Image Harmonization aims to adjust the appearance of the foreground object so that it is compatible with the new composite background. Traditional methods [42, 10, 31, 47, 58, 63] typically rely on adjusting the appearance of the foreground to match with the color statistics of the background. Sunkavalli *et al*. [47] propose to first transfer the visual appearance of the target image to the source image via image histogram matching, and then use alpha blending to produce the composite image. Xue *et al*. [58] suggest to match zones of the (instead complete) histogram is more effective, and propose the zone selection classifier for matching. Reinhard *et al*. [42] propose a color transfer method to match the global color statistics between the source and target images. Lalonde and Efros [31] divide the source and target images into corresponding cluster pairs, and perform the color transfer [42] for each cluster pair locally.

In recent years, many deep methods are proposed for image harmonization. Zhu *et al*. [63] propose to train a CNN classifier to distinguish between real and composite images, and use the learned model to adjust the brightness and contrast model for image composition. Tsai *et al*. [48] present the first end-to-end CNN-based harmonization method to leverage the semantic information of a scene parsing branch to help boost the performance of the harmonization branch. Cun and Pun [13] propose to use spatial attention modules to learn regional appearance changes for harmonization. Chen and Kae [8] propose a GAN-based method to harmonize images with geometric and color consistency constraints. Wu *et al*. [53] propose a GAN-based method to explore both gradient and color constraints for image harmonization. Cong *et al*. [12] construct a large-scale dataset, iHarmony4, and propose a domain verification discriminator to guide the generator to translate the foreground object to the background domain. Ling *et al*. [33] formulate the image harmonization task as a style transfer problem, and propose the region-aware adaptive instance normalization module to model the background style and apply it to the foreground. Guo *et al*. [21] propose to decompose the composite image into reflectance and illumination, and harmonize the two via material consistency and light transfer. Self-supervised learning and transformers have also been applied to image harmonization in [27] and [20], respectively. Methods are also proposed for high-resolution image harmonization [11, 32, 29, 57], multiple objects harmonization [44] and interactive portrait harmonization [49].

Unlike existing methods that harmonize colors in the low dynamic range image domain, we propose to model and harmonize these object colors by tracing back to the high dynamic range scene radiance field.

Image Enhancement aims to produce visually pleasing images with vivid colors and details from low-quality input images of over- or under-exposures. Some meth-

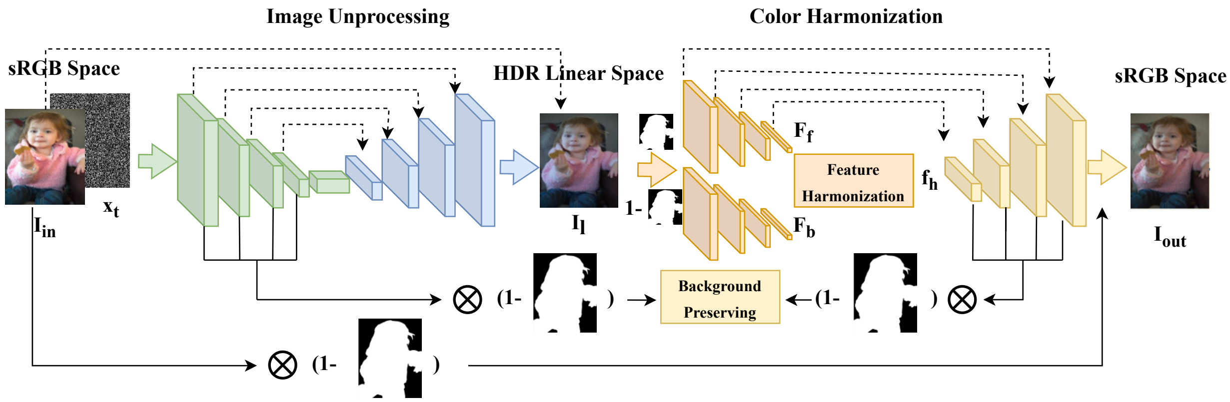


Figure 2. Overview of the proposed method. Given an input composite image, we first convert it into an intermediate linear color space via the image unprocessing process. The Color Harmonization process then harmonizes the foreground colors in the feature domain, and renders the harmonized colors to produce an output sRGB image with the guidance of the background preserving process.

ods [19, 17, 5, 51, 45, 36, 54] rely on the retinex theory to decompose the input image into reflectance and illumination layers, and enhance the illumination layer. Cai *et al*. [6] separately model illuminance and detail layers from multi-exposed images to enhance an under-exposed image. Xu *et al*. [56] decompose and enhance under-exposed images based on frequency information. Moran *et al*. [38] learn a set of local parametric filters for image enhancement. Some methods directly learn an image-to-image mapping using high dynamic range information [18, 59, 46] or adversarial learning [25, 9, 26, 43]. Mahmoud *et al*. [2] propose a coarse-to-fine network to learn color and detail enhancement for addressing over- or under-exposure. Recently, Wang *et al*. [50] build a local color distributions pyramid with a dual-illumination estimation method to handle images of both over-/under-exposures.

While sharing some similarities with image harmonization methods, *e.g*., pixel-wise curve modeling and retinex-based image decomposition, image enhancement methods do not model the discrepancy between source and target scenes. Directly enhancing composite images with image enhancement methods tends to amplify scene discrepancies.

3. Proposed Method

We propose to harmonize cut-and-paste object colors by tracing back them to the scene radiance field. To this end, we propose an image unprocessing method to transfer the input composite image into a linear high dynamic range color space, and a harmonization method to re-render the target object colors by querying the scene radiance information. Figure 2 shows the whole harmonization process.

3.1. Image Unprocessing

Converting camera output sRGB images back to camera raw images, *i.e*., image unprocessing [60, 39, 40, 3, 55, 41], requires a systematic modelling of the ISP operations. Existing methods are typically sensor-specific, requiring additional camera information to convert each image. We note that most camera ISPs typically apply a set of camera-variant linear operations (*e.g*., white balance) to convert CCD data into a camera-independent color space, and then apply another set of non-linear operations (*e.g*., quantization and local enhancement) to render images [28]. Hence, converting the sRGB images back into the intermediate camera-independent color space has two advantages for harmonization. First, colors in this intermediate space response to the scene radiance linearly, which helps recover scene discrepancy for harmonization. Second, it avoids the need to model camera-dependent operations (*e.g*., camera response curves selection [16] or estimation [46]). We formulate a generative diffusion model to address this dynamic range expansion problem of image unprocessing. Although single-image reverse tone mapping is challenging (as it needs to generate missing info in the over-/under-exposed regions), learning such image unprocessing in our task is feasible, as images to be harmonized are typically captured with proper exposures.

Model Formulation. A diffusion model has a forward diffusion process and a reverse denoising process (used for generation). Given a distribution $q\ (x_0)$, the forward diffusion process q is a Markovian noising process [23], which gradually adds noise to x_0 to obtain $x_{1:T}$. Specifically, at each step t, the diffusion process adds random Gaussian noise with a β_t-controlled variance:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \quad (1)$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbb{I}), \quad (2)$$

where $\beta_t \in (0,1), t = 1, ...T$. With the reparameterization trick [30], we sample x_t from each time step t in a closed

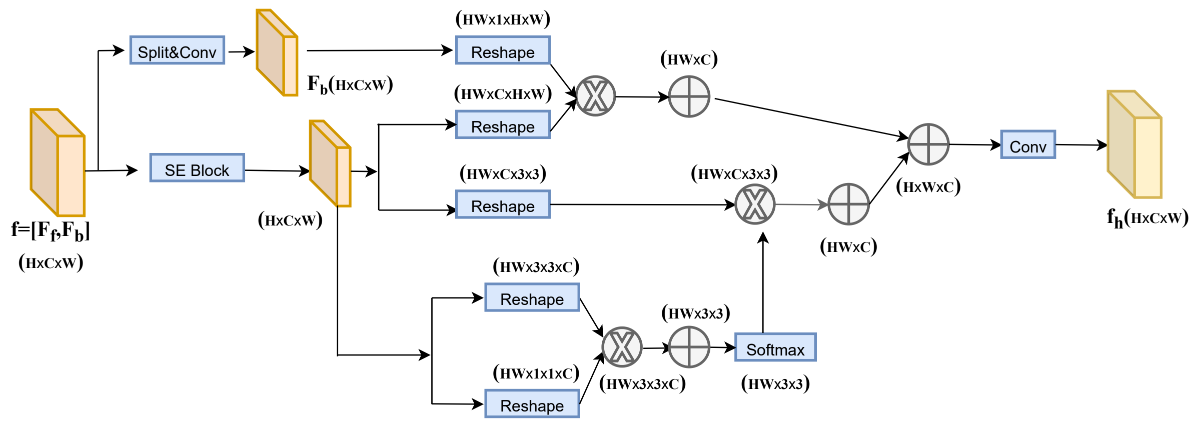


Figure 3. The feature harmonization process. Given the concatenated linear foreground and background features $f = [F_f, F_b]$, it first uses an SE block [24] to adjust the attention for each channel to be the same, and then propagates channel consistency to the spatial domain to produce harmonized foreground features f_h.

form: $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, \mathbb{I})$:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t)\mathbb{I}), \quad (3)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. In this way, we directly derive x_t from $q(x_t|x_0)$ without repeatedly applying the Markovian process q and calculating $q(x_t|x_{t-1})$.

We formulate image unprocessing as a reverse denoising process (*i.e.*, generation process) that is conditioned on the sRGB input image I_{in}. We compute x_0 (*i.e.*, the linear color image I_l) via the reverse diffusion process $p_\theta(x_{t-1}|x_t)$ parameterized by θ with a random Gaussian distribution, *i.e.*, $x_T \sim \mathcal{N}(0, \mathbb{I})$, as:

$$p_\theta(x_{0:T}|I_{in}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, I_{in}), \quad (4)$$

$$p_\theta(x_{t-1}|x_t, I_{in}) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, I_{in}, t), \textstyle\sum_\theta(x_t, I_{in}, t)). \quad (5)$$

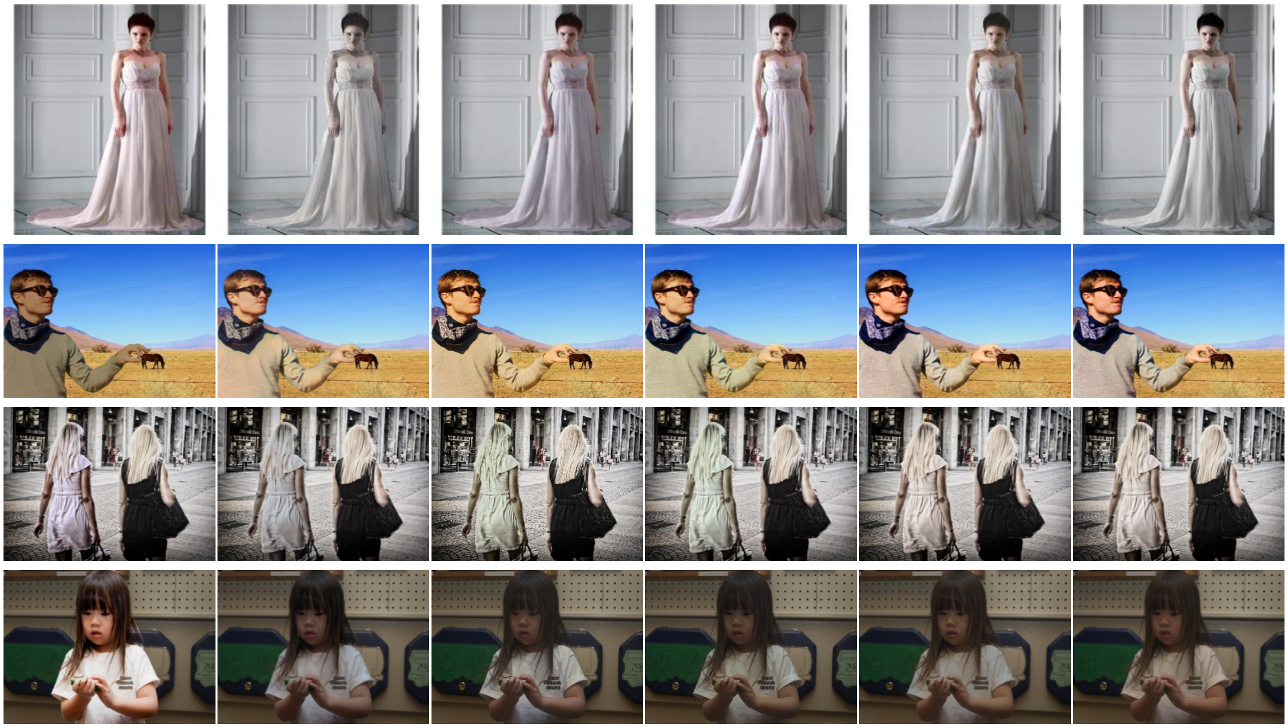
Unlike previous methods that are conditioned on the class labels [14] or shape latents [35], we condition the diffusion process on the sRGB image I_{in} pixel-wisely.

Model Architecture. Our image unprocessing model adopts a fully convolutional encoder-decoder network [34], as shown in Figure 2(left). Given the sampled noise with the sRGB image as the condition, we encode them into a low dimensional latent representation, which is then decoded to reconstruct the linear color image. We leverage the generation ability of reversing the denoising process to expand the dynamic range, and use the encoder-decoder architecture to perform image unprocessing. To facilitate the learning process, we add a skip connection directly from the input image to the output. Instead of learning to generate the linear color image of the whole dynamic range, the image unprocessing network only needs to generate the difference between the input and the output, resulting in a fast reverse diffusion process. We train this network from scratch and use batch normalization and GELU activation for all the convolutional layers. We use the weighted variational bound [23] to optimize this model.

3.2. Harmonization and Rendering

We propose a harmonization module to harmonize the colors of the target object by querying background radiance information and re-render the harmonized image in the sRGB space.

Harmonization in Linear Space. As shown in Figure 2(right), given the reconstructed linear color image I_l, we first separately obtain foreground and background features (*i.e.*, F_f and F_b) via two separate encoders and the foreground mask M, as $F_f = Enc_f(I_l * M)$ and $F_b = Enc_b(I_l * (1 - M))$. Our goal is to harmonize the foreground features F_f based on F_b to obtain $\hat{F}_f$, and then render the harmonized foreground features $\hat{F}_f$ into the sRGB space $\hat{F}_f \rightarrow \hat{I}_f$ conditioned on the rendering process of $F_b \rightarrow I_b$. To this end, we harmonize the concatenated features $f = [F_f, F_b]$ in both channel and spatial dimensions. We implement the channel harmonization by using the squeeze-and-excitation operation [24] to assign consistent attention for each channel of $f = [F_f, F_b]$. As the reweighed features are computed channel-wisely according to both foreground and background representations, the consistent attention indicates that these representations are harmonized to be consistent as well. We then propagate channel harmonization to the spatial domain via the bilat-



(a) Input (b) IIH [21] (c) CDT [11] (d) WBH [29] (e) Ours (f) Ground Truth

Figure 4. Visual comparison on composite images with humans. The proposed method is able to harmonize the colors in the cut-and-paste regions and produce realistic images compared with the state-of-the-art methods.

eral propagation activation functions:

$$y_i^s = \frac{1}{C(f)} \sum_{j \in s} g(\|j - i\|) f_j, \quad (6)$$

$$y_i^r = \frac{1}{C(f)} \sum_{j \in v} h(f_i, f_j) f_j, \quad (7)$$

$$y_i = c(y_i^s, y_i^r), \quad (8)$$

where f_i is the feature channel at position i of the input features f. f_j is a neighboring feature channel around i at position j. y_i^s and y_i^r are the features after spatial and range similarity measurements. The normalization factor is set to $C(f) = N$, where N is the number of positions in f. c represents the pixel-wise summation and a linear transformation of y_i^s and y_i^r via a 1×1 convolutional layer. The bilateral propagation extends the consistency of feature channels to both spatial and range dimensions. In *spatial propagation*, we set the neighboring region s to be of the same spatial resolution as the input features for global propagation. A Gaussian function $g(\cdot)$ [52] is used to compute the spatial contributions from neighboring background features. In *range propagation*, we measure the similarity between features f_i and f_j via $h(\cdot)$ within a neighboring region v around i. The size of v is set to 3×3. The range similarity is computed via the pairwise function $h(\cdot)$ with a dot product operation:

$$h(f_i, f_j) = (f_i)^T (f_j). \quad (9)$$

In this way, the bilateral propagation process harmonizes the foreground colors by considering both global continuity via y_i^s and local consistency via y_i^r. Figure 3 shows the feature harmonization process.

Our bilateral propagation is close in spirit to the non-local block [52], in that for each i, $\frac{1}{C(f)} \sum g(f_i, f_j)$ computes the softmax scores along dimension j. The main difference is the regions of propagation. The non-local block uses feature channels from all positions to generate y_i and the similarity is only measured between f_i and f_j. In contrast, our method considers channel similarity, long-range and neighboring spatial correlations between f_i and f_j for feature harmonization. During the long-range correlation modeling, we query the global scene radiance information using region s of the original background features (*i.e.*, before the channel-wise harmonization), while during the neighboring correlation modeling, we query the local scene information using the background features after channel-wise adjustment. The foreground colors are then set to be consistent to the background both globally and locally.

Background Preserving Guided Rendering. After the harmonization process in the linear embedding space, we assign a decoder to re-project the harmonized linear image

(a) Input (b) IIH [21] (c) CDT [11] (d) WBH [29] (e) Ours (f) Ground Truth

Figure 5. Visual comparison on composite images with general objects. The proposed method is able to harmonize the colors in the cut-and-paste regions and produce more realistic images compared with the state-of-the-art methods.

back to the sRGB space. To avoid further artifacts produced during the foreground rendering process, we leverage the identity property of the background rendering process (*i.e.*, the background should remain identical after the colormetric conversion, harmonization and rendering processes) as the guidance. Since our encoder of the image unprocessing and the color rendering decoder are symmetric, we add a foreground feature consistency constraint to guide the rendering process. In our implementation, we add an L_1 and a Cosine similarity terms to align the foreground features in magnitude and directions, of the image unprocessing encoder and the Color Harmonization decoder (Figure 2 right). We follow previous methods to produce the final harmonized image I_h as:

$$I_{out} = I_o \times M + I_{in} \times (1 - M), \tag{10}$$

where I_o is the decoder output of the Color Harmonization process. We use standard L_1 loss to optimize this model.

3.3. Implementation Details

We have implemented the proposed model under Pytorch, and tested it on a PC with an i7 4GHz CPU and a GTX4090 GPU. The network parameters are initialized using the truncated normal initializer. For loss minimization, we adopt the AdamW optimizer. We first train the image unprocessing network on the Adobe5K dataset [4] for 300 epochs with an initial learning rate of $1e^{-4}$, which is divided by 2 every 75 epochs. We then freeze the image unprocessing network and train the Color Harmonization network on the iHarmony4 dataset [12] for 75 epochs, with an initial learning rate of $1e^{-4}$, which is divided by 10 at the 30th epoch. T in Eq. 1 is empirically set to 4000. It takes around 50 hours to train our model and $0.87s$ for testing a 256×256 image ($3.78s$ for a 1024×1280 image).

4. Results

Evaluation Methods. We compare our method to 7 latest state-of-the-art deep harmonization methods: Dove [12], S^2AM [13], IIH [21], IHT [20], CDT [11], WBH [29] and SCSCo [22]. Since SCSCo [22] does not provide code and results, we directly copy their performances reported in their paper for references. For other methods, we use either the pre-trained models released by their authors or their released results for evaluation. Among them, S^2AM [13] and WBH [29] learn either implicit mapping curves or explicit filters to adjust the appearances of the foreground objects; Dove [12] and SCSCo [22] perform harmonization based on style transfer learning. IIH [21] is retinex-based. It harmonizes the foreground objects in the intermediate illumination layer. CDT [11] and IHT [20] model the harmonization process with pixel-to-pixel mapping while an additional color-to-color mapping is used in CDT [11]. We compare our method to these methods to demonstrate the effectiveness of harmonizing foreground objects in the in-

Datasets	Metrics	S^2AM [13]	Dove [12]	IIH [21]	IHT [20]	CDT [11]	WBH [29]	SCSCo [22]	Ours
HAdobe5K	PSNR↑	33.77	34.34	35.20	36.10	38.24	37.64	38.29	**38.93**
	MSE↓	63.40	52.32	43.02	47.96	20.62	21.89	21.01	**20.11**
	fMSE↓	404.62	380.39	284.21	321.14	-	170.05	165.48	**154.82**
HFlickr	PSNR↑	30.03	29.75	31.34	32.37	33.55	33.63	34.22	**34.76**
	MSE↓	143.45	145.21	105.13	88.41	68.61	64.81	55.83	**54.20**
	fMSE↓	785.65	827.03	716.60	617.26	-	434.06	393.72	**386.12**
HCOCO	PSNR↑	35.47	35.83	37.16	37.87	39.15	38.77	39.88	**39.94**
	MSE↓	41.07	36.72	24.92	20.99	16.25	17.34	13.58	**11.27**
	fMSE↓	542.06	551.01	416.38	377.11	-	298.42	245.54	**217.55**
Hday2night	PSNR↑	34.50	35.53	35.96	36.38	37.95	37.56	37.83	**38.42**
	MSE↓	76.61	56.92	55.53	58.14	36.72	33.14	41.75	**39.79**
	fMSE↓	989.07	1075.71	797.04	823.68	-	542.07	606.80	**587.44**
Average	PSNR↑	34.35	34.75	35.90	36.71	38.23	37.84	38.75	**39.36**
	MSE↓	59.67	52.36	38.71	37.07	23.75	24.26	21.33	**20.77**
	fMSE↓	594.67	532.62	400.29	395.66	-	280.51	248.86	**238.19**

Table 1. Quantitative comparison between the proposed method and state-of-the-art deep harmonization methods on the iHarmony4 dataset [12] at 256×256 image resolution. It shows that the proposed method outperforms existing image harmonization methods. Best performances are marked in **bold**.

termediate linear color space.

In addition, to verify whether our image unprocessing module produces faithful linear images, we perform an internal analysis and compare it to 4 representative state-of-the-art deep networks, including Unprocess [3], HDR-CNN [15], CycleISP [61], and CIE-XYZNet [1]. Among them, Unprocess [3] is a systematic pipeline that converts sRGB images to raw images via a sequence of reverse ISP operations. HDRCNN [15] uses an encoder-decoder network to convert sRGB images into the HDR domain. CycleISP [61] and CIE-XYZNet [1] learn cycle mappings (sRGB-to-raw and raw-to-sRGB).

Evaluation Datasets and Metrics. We follow existing methods to evaluate the harmonization performance by using the Mean Square Error (MSE), foreground MSE (fMSE) and Peak Signal-to-Noise Ratio (PSNR), on the iHarmony4 dataset [12]. When internally analyzing the proposed image unprocessing module on the Adobe5K dataset [4], in addition to the PSNR metric, we also use the widely adopted HDR-VDP-2 [37] metric to measure the image quality based on human perceptions.

4.1. Comparing to State-of-the-art Methods

We compare the proposed method with state-of-the-art image harmonization methods on the standard benchmarks.

Visual Comparison. Figure 4 shows visual comparison where the the cut-and-paste regions contain humans. While the latest methods are effective in adjusting the brightness of the foreground targets to fit the background illumination conditions, they are not able to produce visually pleasing colors, as shown in Figure 4(b-d). In contrast, our method is able to render realistic colors in the cut-and-paste regions, as shown in Figure 4(e). Figure 5 shows some examples where the cut-and-paste regions contain general objects. While existing methods may produce pale colors Figure 5(b,c) or color artifacts Figure 5(d). In contrast, our method produces foreground colors that are more realistic and consistent with the background, as shown in Figure 5(e).

Quantitative Comparison. In addition to the visual evaluation, we also provide quantitative comparison between the proposed method and existing harmonization methods. We first follow existing methods to evaluate the harmonization performance on images of resolution 256×256 in the iHarmony4 benchmark [12]. Table 1 shows the results. We can see that the proposed method outperforms existing methods on all evaluation metrics under all subsets of iHarmony4. Table 2 shows additional comparisons at the original image resolution. Note that CDT [11] only releases their low-resolution results. The comparison shows that our method can handle images of higher resolutions.

4.2. Internal Analysis

As our method first converts the input composite image into the intermediate linear color space via the image unprocessing process, we demonstrate its effectiveness in producing faithful colors in the high dynamic range domain. We evaluate it on the CIE xyz version of the Adobe 5K dataset using the PSNR and HDR-VDP2 metrics. The HDR-VDP2 metric produces a Q score for each test image via a Mean-Opinion-score metric. The Q score indicates the degree of image quality degradation. Table 3 reports the average PSNR and Q score on the test set. The results show that the proposed image unprocessing process can produce more faithful images due to its generation ability.

We now perform ablation studies on our network de-

Datasets	Metrics	IHT [20]	WBH [29]	Ours
HAdobe5K	PSNR↑	33.63	37.80	**38.42**
	MSE↓	56.90	24.37	**23.76**
HFlickr	PSNR↑	29.59	33.37	**33.97**
	MSE↓	135.49	69.19	**61.70**
HCOCO	PSNR↑	34.19	37.69	**38.22**
	MSE↓	44.95	20.93	**20.08**
Hday2night	PSNR↑	35.71	37.15	**37.64**
	MSE↓	63.26	37.28	**30.11**
Average	PSNR↑	33.54	37.23	**37.92**
	MSE↓	58.89	27.62	**23.90**

Table 2. Quantitative comparisons on the iHarmony4 dataset [12] at the original resolution. It shows that our method outperforms all existing image harmonization methods. Best performances are marked in **bold**.

Methods	PSNR↑	Q score↑
Unprocess [3]	22.19	50.33
HDRCNN [15]	27.74	55.61
CycleISP [61]	28.29	54.74
CIE-XYZNet [1]	29.66	56.08
CycleGAN [64]	27.64	56.17
Ours	**30.34**	**58.14**

Table 3. Internal Analysis. We quantitatively compare the converted linear images with existing representative methods using PSNR and Q score. Best performances are marked in **bold**.

Methods	PSNR↑	MSE↓
w/o CC	37.79	24.26
DP→CP	38.28	23.60
w/o FH	38.10	23.94
w/o BP	39.17	21.10
Single Encoder	39.02	22.28
w/o SE Block	38.47	23.79
w/o Spatial Attention	38.19	23.57
Ours	**39.36**	**20.77**

Table 4. Ablation Study of network design on iHarmony4 at 256×256 image resolution. Best performances are marked in **bold**.

sign. We follow previous methods to perform it on the iHarmony4 dataset. In particular, we investigate the following ablated network architectures: (1) we remove the image unprocessing process and directly train a Color Harmonization network (denoted as "w/o CC"); (2) we replace the diffusion process of the image unprocessing with a standard convolution process (denoted as "DP→CP"); (3) we remove the Feature Harmonization from the Color Harmonization process (denoted as "w/o FH"); (4) we remove the Background Preserving of the Color Rendering process (denoted as "w/o BP"); (5) we use a single encoder in the Color Harmonization network ("denoted as Single Encoder"); (6) we remove the SE block from the feature harmonization process ("denoted as w/o SE Block"); and (7) we remove the spatial attention from the feature harmonization process ("denoted as w/o Spatial Attention"). Table 4 reports the performance. It shows that our designs of image unprocessing, Feature Harmonization and background preserving are able to improve the image harmonization performances. We have tried to train DoveNet [12] from scratch in linear space and it yields slightly better results (PSNR:34.94 (+0.19)) but fine-tuning using their pre-trained model degrades the performance (PSNR:33.18 (-1.57)). This is due to the discrepancy between linear and non-linear images, which demonstrates that it is necessary to design a specific image unprocessing model for harmonization. We have also tried ControlNet [62] for harmonization, which, however, does not perform well (PSNR/MAE: 14.36/109.40). As image harmonization requires the harmonized images to be photorealistic, directly using the diffusion-based model may generate visually pleasing but fake image details.

(a) Input (b) Ours (c) Ground Truth

Figure 6. Our method may fail to harmonize the color tones of the inserted target when the new background does not provide sufficient scene illumination information.

5. Conclusion

This paper presents a novel image harmonization method that performs the color harmonization process in the linear color space. Our method includes a novel image unprocessing process to convert the cut-and-paste image into the high dynamic range linear color space, and a novel color harmonization process to harmonize object colors by querying background radiance information. We have conducted extensive experiments on the benchmark datasets to analyze the properties of the proposed method, and shown that it outperforms the state-of-the-art harmonization methods.

Our method does have limitations. It may fail when the background does not contain sufficient scene illumination information. Figure 6 shows such an example, in which the background is completely dark and our method fails to harmonize the color tones of the foreground.

Acknowledgement. This project was partially supported by a GRF grant from the Research Grants Council of Hong Kong (Project No. CityU 11205620), by City University of Hong Kong (9678131), and by the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11214620).

References

[1] Mahmoud Afifi, Abdelrahman Abdelhamed, Abdullah Abuolaim, Abhijith Punnappurath, and Michael S Brown. Cie xyz net: Unprocessing images for low-level computer vision tasks. *IEEE TPAMI*, 2021. 7, 8

[2] Mahmoud Afifi, Konstantinos G Derpanis, Bjorn Ommer, and Michael S Brown. Learning multi-scale photo exposure correction. In *CVPR*, 2021. 3

[3] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *CVPR*, 2019. 3, 7, 8

[4] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR*, 2011. 6, 7

[5] Bolun Cai, Xianming Xu, Kailing Guo, Kui Jia, Bin Hu, and Dacheng Tao. A joint intrinsic-extrinsic prior model for retinex. In *ICCV*, 2017. 3

[6] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE TIP*, 2018. 3

[7] Patrick Cavanagh. The artist as neuroscientist. *Nature*, 2005. 1

[8] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *CVPR*, 2019. 1, 2

[9] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *CVPR*, 2018. 3

[10] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. Color harmonization. *ACM TOG*, 2006. 1, 2

[11] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *CVPR*, 2022. 1, 2, 5, 6, 7

[12] 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, 6, 7, 8

[13] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE TIP*, 2020. 1, 2, 6, 7

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 4

[15] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM TOG*, 2017. 7, 8

[16] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM TOG*, 2017. 3

[17] Xueyang Fu, Delu Zeng, Yue Huang, Xiaoping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *CVPR*, 2016. 3

[18] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *TOG*, 2017. 3

[19] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 2017. 3

[20] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *ICCV*, 2021. 1, 2, 6, 7, 8

[21] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *CVPR*, 2021. 1, 2, 5, 6, 7

[22] Yucheng Hang, Bin Xia, Wenming Yang, and Qingmin Liao. Scs-co: Self-consistent style contrastive learning for image harmonization. In *CVPR*, 2022. 6, 7

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 4

[24] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 4

[25] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. DSLR-quality photos on mobile devices with deep convolutional networks. In *ICCV*, 2017. 3

[26] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE TIP*, 2021. 3

[27] 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. 1, 2

[28] Hakki Can Karaimer and Michael Brown. A software platform for manipulating the camera imaging pipeline. In *ECCV*, 2016. 3

[29] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *ECCV*, 2022. 1, 2, 5, 6, 7, 8

[30] Diederik Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 3

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

[32] Jingtang Liang, Xiaodong Cun, and Chi-Man Pun. Spatial-separated curve rendering network for efficient and high-resolution image harmonization. In *ECCV*, 2022. 2

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

[34] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 4

[35] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, 2021. 4

[36] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, 2022. 3

[37] Rafat Mantiuk, Kim Joong, Allan Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM TOG*, 2011. 7

[38] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory G. Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *CVPR*, 2020. 3

[39] Rang Nguyen and Michael Brown. Raw image reconstruction using a self-contained srgb-jpeg image with only 64 kb overhead. In *CVPR*, 2016. 3

[40] Rang Nguyen and Michael Brown. Raw image reconstruction using a self-contained srgb–jpeg image with small memory overhead. *IJCV*, 2018. 3

[41] Hao Ouyang, Zifan Shi, Chenyang Lei, Ka Lung Law, and Qifeng Chen. Neural camera simulators. In *CVPR*, 2021. 3

[42] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE CGA*, 2001. 1, 2

[43] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Minghsuan Yang. Low-light image enhancement via a deep hybrid network. *IEEE TIP*, 2019. 3

[44] Xuqian Ren and Yifan Liu. Semantic-guided multi-mask image harmonization. In *ECCV*, 2022. 2

[45] Liu Risheng, Ma Long, Zhang Jiaao, Fan Xin, and Luo Zhongxuan. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *CVPR*, 2021. 3

[46] Aashish Sharma and Robby T Tan. Nighttime visibility enhancement by increasing the dynamic range and suppression of light effects. In *CVPR*, 2021. 3

[47] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM TOG*, 2010. 1, 2

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

[49] Jeya Maria Jose Valanarasu, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Jose Echevarria, Yinglan Ma, Zijun Wei, Kalyan Sunkavalli, and Vishal Patel. Interactive portrait harmonization. *arXiv:2203.08216*, 2022. 2

[50] Haoyuan Wang, Ke Xu, and Rynson W.H. Lau. Local color distributions prior for image enhancement. In *ECCV*, 2022. 3

[51] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, 2019. 3

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

[53] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In *ACM MM*, 2019. 1, 2

[54] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *CVPR*, 2022. 3

[55] Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *CVPR*, 2021. 3

[56] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *CVPR*, 2020. 3

[57] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *ECCV*. 2

[58] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM TOG*, 2012. 1, 2

[59] Xin Yang, Ke Xu, Yibing Song, Qiang Zhang, Xiaopeng Wei, and Rynson Lau. Image correction via deep reciprocating HDR transformation. In *CVPR*, 2018. 3

[60] Lu Yuan and Jian Sun. High quality image reconstruction from raw and jpeg image pair. In *ICCV*, 2011. 3

[61] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *CVPR*, 2020. 7, 8

[62] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 8

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

[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *ICCV*, 2017. 8