Intrinsic Image Harmonization

Zonghui Guo[1] Haiyong Zheng[1,*] Yufeng Jiang[1] Zhaorui Gu[1] Bing Zheng[1,2]
[1]Underwater Vision Lab (http://ouc.ai), Ocean University of China
[2]Sanya Oceanographic Institution, Ocean University of China
{guozonghui,jiangyufeng7526}@stu.ouc.edu.cn, {zhenghaiyong,guzhaorui,bingzh}@ouc.edu.cn

Abstract

Compositing an image usually inevitably suffers from inharmony problem that is mainly caused by incompatibility of foreground and background from two different images with distinct surfaces and lights, corresponding to material-dependent and light-dependent characteristics, namely, reflectance and illumination intrinsic images, respectively. Therefore, we seek to solve image harmonization via separable harmonization of reflectance and illumination, i.e., intrinsic image harmonization. *Our method is based on an autoencoder that disentangles composite image into reflectance and illumination for further separate harmonization. Specifically, we harmonize reflectance through material-consistency penalty, while harmonize illumination by learning and transferring light from background to foreground, moreover, we model patch relations between foreground and background of composite images in an inharmony-free learning way, to adaptively guide our intrinsic image harmonization. Both extensive experiments and ablation studies demonstrate the power of our method as well as the efficacy of each component. We also contribute a new challenging dataset for benchmarking illumination harmonization. Code and dataset are at https://github.com/zhenglab/IntrinsicHarmony.*

1. Introduction

The visual appearance of two images will be distinct due to different light and scene while imaging [54, 60]. Thus, compositing an image, *i.e.*, extracting a foreground region in one image and pasting it with the background of another image, will inevitably suffer from the inharmony problem caused by distinct appearance between the two images (see Figure 1 for example), which significantly degrades the quality of composite result [49, 12, 11]. Besides, many computer vision tasks, especially image/video synthesis, such as image editing [38, 2, 44], image inpainting [34, 55, 41], and image stitching [9, 56, 57], will also encounter this inharmony problem because of the compositing process. However, human visual system is very sensitive to the inharmony in appearance, *e.g.*, human eyes can identify very subtle distinctions in color and contrast [30, 54]. Therefore, image harmonization, which aims to make the appearance of foreground and background in the composite image compatible [47, 49, 11, 12], is full of challenges.

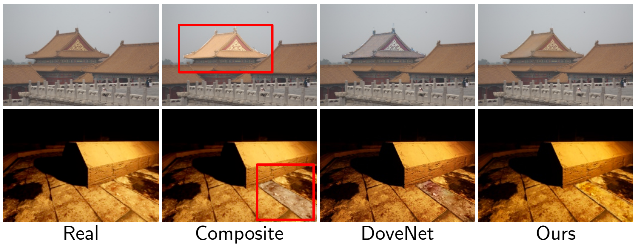

Figure 1. Thanks to separate intrinsic image harmonization, our method can adjust the illumination of foreground to make it compatible with background while keep reflectance constant. We show examples from iHarmony4 [11] (top) and our HVIDIT (bottom).

Essentially, the appearance of a natural image depends on various factors in the scene, such as illumination, material, and shape [58, 3]. Moreover, human visual system is remarkable in its ability to estimate characteristics intrinsic to the scene, such as color, size, shape, or illumination, where each intrinsic characteristic corresponds to one *intrinsic image* [4]. Thus, humans are able to distinguish composite images, mainly owing to the apparent ability to estimate intrinsic characteristics in unfamiliar scenes. Besides, the light intensity values represented in an image actually encode all the characteristics of the corresponding scene points [4]. Hence, harmonizing each intrinsic image separately, rather than adjusting intensity values for harmonization, is essential and crucial for image harmonization.

The inharmony in composite images is caused by the incompatibility of foreground and background from two dif-

[*] Corresponding author (zhenghaiyong@ouc.edu.cn).

This work was supported in part by the Finance Science and Technology Project of Hainan Province under Grant ZDKJ202017, and the National Natural Science Foundation of China under Grant 61771440 and Grant 41776113.

ferent scenes [49, 11], mainly lying in: (1) the boundaries of foreground surfaces (*e.g.*, object, person, *etc.*) are not in harmony with the background [47], corresponding to the intrinsic characteristic of surface; and (2) the illumination of foreground (from one image) and background (from another image) is not harmonious [47, 11], corresponding to the intrinsic characteristic of illumination. Thus, to solve image harmonization, it is intuitive and will be much beneficial to disentangle composite image into intrinsic images of surface and illumination for separate harmonization.

Retinex theory [33, 31] addresses the problem of separating illumination from reflectance in a given image, where illumination represents the intrinsic light-dependent characteristic of scene, while reflectance describes the intrinsic material-dependent characteristic of surface that is invariant to illumination and imaging conditions [26, 15, 35]. Therefore, in this work, we seek to solve image harmonization via separable harmonization of reflectance and illumination intrinsic images, *i.e.*, *intrinsic image harmonization*.

In this paper, we formulate image harmonization as an autoencoder that internally disentangle composite image into reflectance and illumination for separate harmonization. For reflectance, we leverage material-consistency as a penalty cue to harmonize foreground boundaries while keeping reflectance constant. For illumination, we design a lighting strategy to harmonize the incompatibility of foreground and background illumination. Moreover, we note that the inharmony will be more visually obvious if background and foreground have similar material (*e.g.*, the top example in Figure 1), and a single natural image has powerful internal statistics that small patches recur abundantly [18, 61, 22], so compositing an image realistically and meaningfully should also follow this statistics, hence, we devise an inharmony-free learning way to model patch relations of foreground and background from composite images for adaptively guiding intrinsic image harmonization.

Note that, our method is not aimed to achieve physically based intrinsic images, rather our goal is to tackle image harmonization problem via intrinsic image properties. Actually, we are interested in relative, rather than absolute, reflectance and illumination, relying on intrinsic image properties, which aims at harmonizing foreground to be compatible with background in the composite image. Therefore, we leverage both intrinsic image (as motivation) and image harmonization (as objective) to carefully design our network as well as losses, for eliminating inharmony of composite images (but not for intrinsic image decomposition).

Our contributions include: (1) we propose a novel method to harmonize composite images via separately harmonizing reflectance and illumination intrinsic images, as far as we can tell, this is the first to model image harmonization based on intrinsic image theory; (2) we design a lighting strategy to learn light from image and transfer light from background to foreground for illumination harmonization; (3) we devise an implicit way to learn inharmony-free patch relations between foreground and background from composite images for adaptively guiding intrinsic image harmonization; (4) our method achieves state-of-the-art performance on both public synthesized dataset and real composite images, besides, we build a new HVIDIT dataset for benchmarking illumination harmonization.

2. Related Work

Image Harmonization. Early contributions in image harmonization have focused on better matching techniques to ensure consistent appearance of low-level statistics, such as color [40], gradient [38, 48, 24], and hybrid [47]. While then, high-level visual realism of composite images has been taken into account for appearance harmonization [30, 25, 54, 60]. Recently, convolutional neural networks (CNNs) have been developed for end-to-end image harmonization, *i.e.*, Tsai *et al.* [49] used a skip-connected CNN to capture context and semantic information of composite images during harmonization, and Cun *et al.* [12] designed an additional Spatial-Separated Attention Module to learn regional appearance changes in low-level features for harmonization, then Cong *et al.* [11, 10] considered image harmonization as domain translation to transform foreground domain to background domain. Different from all existing methods, we devote to solve harmonization from a novel perspective of intrinsic image harmonization.

Intrinsic Images. The scene in an image can be described in terms of intrinsic characteristics, such as reflectance and illumination, and there is one image for each intrinsic characteristic, all in registration with the original image, namely intrinsic image [4]. Due to its inherent ill-posedness, intrinsic image decomposition from a single image cannot be solved without prior knowledge on reflectance and illumination [4]. Thus, Retinex theory [33, 31] has been proposed to show that reflectance could be separated from illumination if the illumination was assumed to vary smoothly [20]. Then variational methods introduced regularization on reflectance and illumination based on the properties of them [8, 32, 26], such that intrinsic image decomposition could be applied for many computer vision and graphics tasks, such as illumination estimation [45, 17, 50], image enhancement [16, 21, 52, 50], shape from shading [3, 43, 53], and image relighting [39, 59, 36]. Recent advances mainly relied on deep neural networks (DNNs) for the decoupling of reflectance from illumination and achieved good performance [15, 6, 35]. In our work, we build a novel autoencoder-based DNN architecture to harmonize composite images via separable reflectance and illumination intrinsic image harmonization.

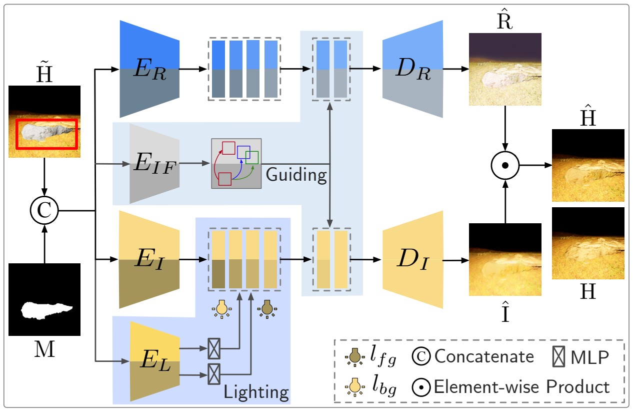

Figure 2. Our autoencoder-based architecture for intrinsic image harmonization, which disentangles composite image into reflectance (via encoder E_R and decoder D_R) and illumination (via encoder E_I and decoder D_I) for separate harmonization. l_{fg} and l_{bg} are foreground and background light latent code extracted by a light learner (encoder E_L) for light transferer shown in Figure 3. Encoder E_{IF} is for inharmony-free patch relation modeling.

3. Method

Given a pair of real image $\mathbf{H}$ and composite image $\tilde{\mathbf{H}}$, with a foreground mask $\mathbf{M}$ that indicates the inharmonious region, our goal is to learn a model $\mathbf{\Phi}$ that receives $\tilde{\mathbf{H}}$ and $\mathbf{M}$ as inputs and produces a harmonized image $\hat{\mathbf{H}}$ as output, where $\hat{\mathbf{H}}$ is expected to be as harmonious as $\mathbf{H}$.

To harmonize $\tilde{\mathbf{H}}$ (to $\hat{\mathbf{H}}$), our method aims to separately harmonize intrinsic images of reflectance $\tilde{\mathbf{R}}$ (to $\hat{\mathbf{R}}$) and illumination $\tilde{\mathbf{I}}$ (to $\hat{\mathbf{I}}$) that are material-dependent and light-dependent respectively. We first build an autoencoder-based architecture to disentangle composite image into reflectance and illumination intrinsic images for separate harmonization, then harmonize reflectance via material-consistency penalty, while harmonize illumination by adjusting foreground illumination to be compatible with background, further we model inharmony-free patch relations between foreground and background of composite images for guiding intrinsic image harmonization. The architecture is illustrated in Figure 2, and we use the mask to separate foreground and background in our lighting and guiding processes.

3.1. Intrinsic Image Harmonization

Decomposition of intrinsic images from a single image is a classic ill-posed problem as information is confounded in the light-intensity image [4, 35]. The only way to decode the confounded information is, apparently, to make assumptions about the world and to exploit the constraints they imply [4]. According to Retinex theory with the assumption of ideal Lambertian surface [33, 20], reflectance is piece-wise constant while illumination is smooth, which allows for the decoupling of a reflectance image corresponding to large image gradients from an illumination image corresponding to small image gradients [26, 15].

So we can disentangle the composite image $\tilde{\mathbf{H}}$ into reflectance and illumination intrinsic images $\tilde{\mathbf{R}}$ and $\tilde{\mathbf{I}}$ by:

$$\tilde{\mathbf{H}} = \tilde{\mathbf{R}} \odot \tilde{\mathbf{I}}, \tag{1}$$

where $\odot$ is element-wise product. Therefore, the harmonization problem of $\hat{\mathbf{H}} = \mathbf{\Phi}(\tilde{\mathbf{H}}) \approx \mathbf{H}$ (where $\hat{\mathbf{H}} = \hat{\mathbf{R}} \odot \hat{\mathbf{I}}$ and $\mathbf{H} = \mathbf{R} \odot \mathbf{I}$) can be divided into two harmonization subproblems of $\hat{\mathbf{R}} \approx \mathbf{R}$ and $\hat{\mathbf{I}} \approx \mathbf{I}$, namely, reflectance harmonization and illumination harmonization, respectively.

Note that most intrinsic image work is only interested in recovering relative reflectance and illumination of a given scene [20]. That is, the estimated reflectance and illumination images are each allowed to be any scalar multiple of the true reflectance and illumination (refer to Equation (1)). Therefore, the reflectance and illumination obtained by our method in this work are also relative.

As we have both composite images and real images to learn from, the learning objective of intrinsic image harmonization is also reconstructive, in other words, the model is trained to disentangle composite image into reflectance and illumination for separate harmonization while the combination of the two harmonized intrinsic images gives back real image. This results in an autoencoding pipeline where we can embed harmonization into the process from composite image decomposition to real image recomposition via loss:

$$\mathcal{L}_{rec} = \mathbb{E}_{(\hat{\mathbf{H}}, \mathbf{H})}\left[\|\hat{\mathbf{H}} - \mathbf{H}\|_1 \right]. \tag{2}$$

Reflectance Harmonization. The reflectance, or albedo, describes how the material of an object reflects light independent of viewpoint and illumination [20]. That is, reflectance is material-dependent yet light-independent. For image harmonization, we assume that the foreground in a composite image is semantically reasonable (or else, it will be out of the scope of harmonization). Thus, the material of foreground in both composite image and real image (as well as harmonized image) should be consistent, giving the constant reflectance constraint, namely, $\tilde{\mathbf{R}} \approx \mathbf{R} \approx \hat{\mathbf{R}}$ (also $\nabla\tilde{\mathbf{R}} \approx \nabla\mathbf{R} \approx \nabla\hat{\mathbf{R}}$, ∇ denotes gradient). In addition, with the assumption that large image gradients correspond to changes in reflectance and the prior that reflectance should be spatially smooth to be a "visually pleasing" image [26], we have the constraints of $\nabla\tilde{\mathbf{H}} \approx \nabla\tilde{\mathbf{R}}$ and $\nabla\mathbf{H} \approx \nabla\mathbf{R}$ for composite and real images respectively. Therefore, we obtain $\nabla\hat{\mathbf{R}} \approx \nabla\mathbf{H}$ as a constraint to harmonize reflectance, yielding reflectance harmonization loss:

$$\mathcal{L}_{RH} = \mathbb{E}_{(\nabla\hat{\mathbf{R}}, \nabla\mathbf{H})}\left[\|\nabla\hat{\mathbf{R}} - \nabla\mathbf{H}\|_1 \right]. \tag{3}$$

This loss actually keeps the reflectance of foreground in composite image to be as close to that in real image as possible, so as to maintain the consistent material. Meanwhile, the foreground boundaries in composite image can also be harmonized to be compatible with the background by leveraging both $\mathcal{L}_{RH}$ and CNNs.

Illumination Harmonization. The illumination accounts for shading effects, including shading due to geometry, shadows and interreflections [20, 3]. So illumination is light-dependent while retaining main structures. Essentially, the main inharmony existing in composite image is caused by incompatible illumination between foreground (from one image) and background (from another image), since they are usually captured under different lighting conditions. Thus, to harmonize illumination, we need to adjust foreground illumination $\tilde{\mathbf{I}}_{fg}$ through background illumination $\tilde{\mathbf{I}}_{bg}$ since $\tilde{\mathbf{I}}_{bg} \approx \mathbf{I}$, such that the illumination of foreground and background will be made compatible. To do this, we design a novel lighting strategy to first learn light and then transfer light from background to foreground (see Section 3.2 for details). Besides, with the assumption of small image gradients corresponding to illumination (*i.e.*, illumination is smooth), we have the constraint of $\nabla\hat{\mathbf{I}} \approx 0$ for decoupling, providing illumination smooth loss:

$$\mathcal{L}_{IS} = \nabla\hat{\mathbf{I}}. \tag{4}$$

Furthermore, since reflectance is restricted to unit interval, we therefore can add the assumption that illumination image is close to its source image (the intensity image) according to Retinex theory (refer to Equation 1, *i.e.*, reflectance tends to white [26]). To better decouple illumination for intrinsic image harmonization, we then make a penalty to force a proximity between harmonized illumination and real image, producing illumination harmonization loss:

$$\mathcal{L}_{IH} = \mathbb{E}_{(\hat{\mathbf{I}},\mathbf{H})}\left[\|\hat{\mathbf{I}} - \mathbf{H}\|_2\right]. \tag{5}$$

Overall, the learning objective for intrinsic image harmonization is given by the combination of reconstruction error as well as reflectance and illumination penalties:

$$\mathcal{L}(\mathbf{\Phi}; \tilde{\mathbf{H}}, \mathbf{M}) = \mathcal{L}_{rec} + \lambda_{RH}\mathcal{L}_{RH} + \lambda_{IS}\mathcal{L}_{IS} + \lambda_{IH}\mathcal{L}_{IH}, \tag{6}$$

where λ_{RH}, λ_{IS}, and λ_{IH} are weighting factors to balance the contribution of different losses.

3.2. Light Learning and Transferring

The inharmony of composite images mainly attributes to incompatible illumination between foreground and background captured under different lighting conditions. Thus, to harmonize this incompatibility, we first extract the light of foreground and background and then transfer the light from background to foreground, by designing a lighting structure with a light learner followed by a light transferer, as shown in Figure 2 (bottom) and 3 (right), respectively.

Benefiting from the autoencoding pipeline conditioned on intrinsic image harmonization (Equation (6)), composite image can be encoded to illumination latent space for recovering illumination intrinsic image (Figure 2). We hence carry out our lighting strategy in this illumination latent space, where an encoder can be served to map an image to its illumination latent representation. Since the illumination latent features are still high-dimensional and embedded with semantic information of the scene, so we further feed them into multi-layer perceptron (MLP) to fetch low-dimensional light latent code as light representation. In this way, we harvest a light learner to extract light from image.

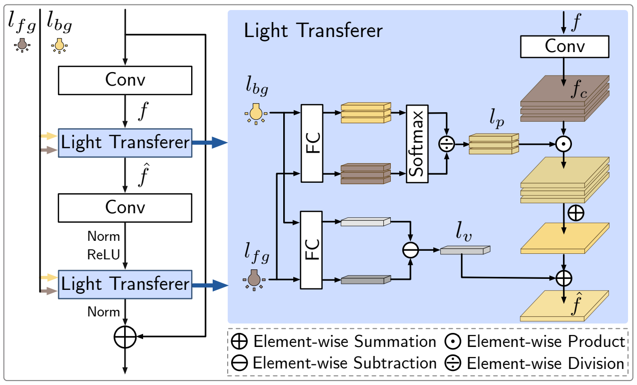

Figure 3. Our lighting ResBlock (left) and light transferer (right) illustration. l_{fg} and l_{bg} are foreground light and background light extracted by a light learner shown in Figure 2 (bottom).

According to the fact that human eye has three different types of color sensitive cones, the response of eye to light is best described in terms of three "tristimulus values" [42, 37]. So we consider to compress light latent code to three elements for representing the light of a scene in image. However, in terms of purely visual phenomena of color derived from reflected light, the three essential traits are hue, value and chroma [46], where hue is the name of a color (pure) while chroma is a property of a color in relation to how pure it is, and value describes how light or dark a color is [1]. Thus, for the sake of simplicity, we represent the light as separate color proportions of three color elements by combining hue with chroma and independent color values, which can be obtained by feeding light latent code into a MLP plus softmax for proportions $l_p \in \mathbb{R}^{3\times C}$ and a MLP for values $l_v \in \mathbb{R}^{1\times C}$ (C is the channel number of feature maps). Then, to transfer light from background to foreground, we adjust the light on illumination latent representation of composite image by:

$$\hat{f} = \sum_{n=1}^{3} f_n \cdot \frac{l_{p_n}^{bg}}{l_{p_n}^{fg}} + (l_v^{bg} - l_v^{fg}), \tag{7}$$

where f and $\hat{f}$ are illumination latent representation before and after light transfer, n denotes the index of color element, bg and fg mean background and foreground respectively. The ratio $\frac{l_p^{bg}}{l_p^{fg}}$ and the difference $l_v^{bg} - l_v^{fg}$ respectively adjust color proportions and color values of foreground light to be close to that of background light. In this way, light transfer takes place. Thus, we construct a novel light transfer layer that can be embedded into various CNN blocks (*e.g.*, ResBlock) for multi-layer light transferring. In this work, we make a lighting ResBlock for our light transferer.

3.3. Inharmony-Free Patch Relation Modeling

As small patches in a single natural image tend to recur abundantly within and across different scales of same image [61], while we also note that similar materials (represented as patches) appeared in both foreground and background of composite image are more prone to be visually inharmonious, so it would be more helpful if we guide intrinsic image harmonization by telling model patch relations on similarity between foreground and background. Solving this issue is closely related to traditional patch matching [2].

However, different from previous applications using patch matching (*e.g.*, image editing and image inpainting [2, 55, 41]), similar patches between foreground and background of composite image are visually different due to inharmony problem, thus it doesn't work to directly match patches on composite image. The only hope of solving this issue is to learn to eliminate the influence of inharmony for patch relation modelling, which we call inharmony-free.

To do this, we provide an implicit way to force an encoder to learn an inharmony-free representation for composite image, yet modeling patch relations between foreground and background of composite image, via the following inharmony-free loss:

$$\mathcal{L}_{IF} = 1 - \mathcal{S}\left[\mathbb{E}_C(\Psi(\tilde{\mathbf{H}})), \mathbf{H}'\right], \tag{8}$$

where $\Psi(\tilde{\mathbf{H}})$ denotes the encoder receiving composite image as input and producing inharmony-free feature maps as output, C is the channel number of $\Psi(\tilde{\mathbf{H}})$, $\mathbf{H}'$ represents the downscaled grayscale real image with the same size of $\Psi(\tilde{\mathbf{H}})$, and $\mathcal{S}$ is the similarity function. We use SSIM [51] to measure structural similarity in this work.

We model patch relations by computing patch covariance of inharmony-free foreground and background feature maps separated from $\Psi(\tilde{\mathbf{H}})$. Note that, although the covariance is computed in terms of pixels, they actually correspond to patches of input due to receptive field of CNN, such that patch relations can be modeled by this inharmony-free learning in an implicit way. We then design a guiding block for bottleneck of autoencoder to reconstruct foreground features, by using patches extracted from background features as filters to deconvolve inharmony-free patch relations. We therefore add this loss with a weighting factor for balance as $\lambda_{IF}\mathcal{L}_{IF}$ to Equation 6 yielding our final total loss:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{RH}\mathcal{L}_{RH} + \lambda_{IS}\mathcal{L}_{IS} + \lambda_{IH}\mathcal{L}_{IH} + \lambda_{IF}\mathcal{L}_{IF}. \tag{9}$$

4. Experiments

4.1. Datasets and Metrics

Public Synthesized Dataset. To evaluate the performance of our method on image harmonization, we conduct experiments on public synthesized iHarmony4 dataset [11] consisting of 4 sub-datasets: HCOCO, HAdobe5k, HFlickr, and Hday2night, each of which contains synthesized composite images, foreground masks of composite images, and corresponding real images. We follow the same settings of iHarmony4 as DoveNet [11] in this work.

Our HVIDIT Dataset. To benchmark image harmonization specifically for illumination, we construct a new synthesized dataset named HVIDIT, which is generated based on VIDIT (Virtual Image Dataset for Illumination Transfer) [13] for Harmonization. VIDIT is used for relighting challenge in the AIM workshop (part of ECCV 2020) [14], including 390 different unreal engine scenes, each captured with 40 illumination settings, among which 300 scenes for training and 90 scenes for testing. We use the publicly available train data in this work since ground-truth test data remain private. Considering the similarity of VIDIT and day2night [28], *i.e.*, the same scene is captured under different lighting conditions, we follow the same way as constructing Hday2night [11] to generate composite images of our HVIDIT, but exclude those cases with obvious shadow since they actually destroy the semantics of background. Thus, we finally obtain 3007 images of 276 scenes for training and 329 images of 24 scenes for testing.

Evaluation Metrics. Following [49, 12, 11], we also use Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity (SSIM) [51] scores on RGB channels as evaluation metrics. Besides, we report foreground MSE (fMSE) [11] and foreground SSIM (fSSIM) that only calculates MSE and SSIM in the foreground region respectively as additional metrics, measuring how well the foreground is harmonized. Noting that, we compute fMSE and fSSIM over each single image and then take average across the dataset, so that they can be regarded as the indicator in evaluating harmonization generalization ability of the method. Whereas, we argue that MSE and SSIM essentially measure the average errors over all pixels across the dataset, thus are not very suitable for tasks like harmonization with many pixels (background) unchanged. In addition, we use small window size (11×11) [19] for our SSIM and fSSIM to avoid untrue similarity reflection due to different sizes of foreground for harmonization.

Implementation Details. Reflectance and illumination outputs are normalized to $[0, 1]$ to recover $\hat{\mathbf{H}}$. We train our model using Adam optimizer [27] with parameters of $\beta_1 = 0.5$, $\beta_2 = 0.999$, and learning rate $\alpha = 0.0001$. We set $\lambda_{RH} = 0.1$, $\lambda_{IS} = 0.01$, $\lambda_{IH} = 0.1$, and $\lambda_{IF} = 1$ in experiments. We include more details in *supplementary file*.

Dataset	Metric	Composite	Retinex-Net [52]	DIH [49]	S^2AM [12]	DoveNet [11]	Ours (base)	Ours (base+lighting)	Ours (base+guiding)	Ours
HCOCO	PSNR↑	33.94	33.50	34.69	35.47	35.83	36.24	37.03	36.04	**37.16**
	MSE↓	69.37	69.18	51.85	41.07	36.72	32.36	25.45	30.62	**24.92**
	fMSE↓	996.59	958.16	798.99	542.06	551.01	510.98	429.09	477.73	**416.38**
	SSIM↑	**0.9853**	0.9554	0.9433	0.9417	0.9546	0.9777	0.9801	0.9751	0.9803
	fSSIM↑	0.8257	0.8242	0.8225	0.8477	0.8473	0.8301	0.8587	0.8399	**0.8619**
HAdobe5k	PSNR↑	28.16	27.99	32.28	33.77	34.34	33.90	35.17	34.16	**35.20**
	MSE↓	345.54	335.27	92.65	63.40	52.32	57.14	44.15	48.64	**43.02**
	fMSE↓	2051.61	1961.14	593.03	404.62	380.39	381.39	284.25	340.18	**284.21**
	SSIM↑	**0.9483**	0.8716	0.8723	0.8884	0.8841	0.9297	0.9354	0.9354	0.9356
	fSSIM↑	0.7294	0.7254	0.7777	0.8120	0.8309	0.7989	0.8357	0.8031	**0.8364**
HFlickr	PSNR↑	28.32	28.06	29.55	30.03	30.21	30.64	31.29	30.69	**31.34**
	MSE↓	264.35	262.63	163.38	143.45	133.14	127.95	107.17	118.93	**105.13**
	fMSE↓	1574.37	1534.89	1099.13	785.65	827.03	814.98	728.06	773.59	**716.60**
	SSIM↑	**0.9618**	0.9229	0.9114	0.9120	0.9272	0.9523	0.9587	0.9490	0.9590
	fSSIM↑	0.8031	0.7967	0.7984	0.8233	0.8235	0.7994	0.8280	0.8031	**0.8297**
Hday2night	PSNR↑	34.01	33.15	34.62	34.50	35.27	34.51	35.89	33.87	**35.96**
	MSE↓	109.65	109.45	82.34	76.61	**51.95**	80.30	57.51	104.09	55.53
	fMSE↓	1409.98	1365.01	1129.40	989.07	1075.71	1160.08	871.09	1082.51	**797.04**
	SSIM↑	**0.9606**	0.8942	0.8838	0.8775	0.8961	0.9239	0.9308	0.9285	0.9302
	fSSIM↑	0.6353	0.6365	0.6277	0.6374	0.6194	0.5923	0.6448	0.6039	**0.6449**
All	PSNR↑	31.63	31.28	33.41	34.35	34.76	34.90	35.82	34.86	**35.90**
	MSE↓	172.47	169.16	76.77	59.67	52.33	51.14	40.62	47.07	**38.71**
	fMSE↓	1376.42	1322.57	773.18	594.67	532.62	518.83	428.21	484.62	**400.29**
	SSIM↑	**0.9714**	0.9262	0.9179	0.9217	0.9299	0.9599	0.9638	0.9598	0.9699
	fSSIM↑	0.7917	0.7889	0.8032	0.8308	0.8357	0.8133	0.8447	0.8208	**0.8469**

Note: we refer to the results reported in paper [11] or train the models to obtain the unavailable results for comparison.

Table 1. Quantitative comparison across four sub-datasets of iHarmony4 [11]. The ↑ indicates the higher the better, and ↓ indicates the lower the better. The best results are denoted in boldface. We compute fMSE and fSSIM at image level for better harmonization reflection.

4.2. Comparison with State-of-the-arts

We compare our method with state-of-the-art image harmonization methods: DIH [49], S^2AM [12] and DoveNet [11], as well as an intrinsic image decomposition algorithm: Retinex-Net [52]. We don't compare with traditional image harmonization methods since they have been proven to perform worse than deep learning methods [49, 12, 11]. Besides, although there exists cutting-edge work on intrinsic image decomposition [15, 5, 35], none of them is specifically aimed at image harmonization, also they mainly require additional ground-truth intrinsic images for supervision and focus on how to decompose better, thus it is not suitable to compare with them for harmonization task. Moreover, for ablation study, we build three variants of our intrinsic image harmonization method as follows: base (ours without lighting and guiding), base+lighting (ours without guiding), and base+guiding (ours without lighting).

Table 1 shows quantitative comparison results of image harmonization across four sub-datasets of iHarmony4 [11], as can be seen, (1) even our base model still outperforms other methods across entire dataset, (2) our base model does not perform well on Hday2night possibly due to big lighting difference, (3) either lighting or guiding helps to improve performance and lighting boost with significant gains, and (4) our method with both lighting and guiding achieves state-of-the-art performance.

Figure 4 illustrates the qualitative comparison results of image harmonization on iHarmony4 and HVIDIT datasets, which demonstrate that, thanks to intrinsic image harmonization, our method achieves the best visual effect comparable to real images, even for very difficult cases such as the fourth row of Figure 4 with very small foreground. More results are shown in *supplementary file* for further reference.

Furthermore, we conduct an additional experiment by inverting the normal masks, that is, exchanging foreground and background to yield inverted masks, so that our method tries to harmonize the background according to the foreground. Figure 5 shows harmonized results with normal masks (middle row) and inverted masks (bottom row) for contrast, indicating that our method can produce promising harmonized outputs from arbitrary foreground masks.

4.3. Analysis of Lighting

We then conduct to analyze efficacy of our light learning and transferring (lighting) as follows: (1) our method without lighting and guiding (base), (2) replace our light transferer with AdaIN [23] (base+AdaIN), (3) base with lighting but only background light (with l^{bg} only), and (4) base with lighting. The quantitative comparison in Table 2 shows that, AdaIN, aiming to transfer style information, has some positive effect on performance, but not as much as our lighting, also background light is more important for harmonization.

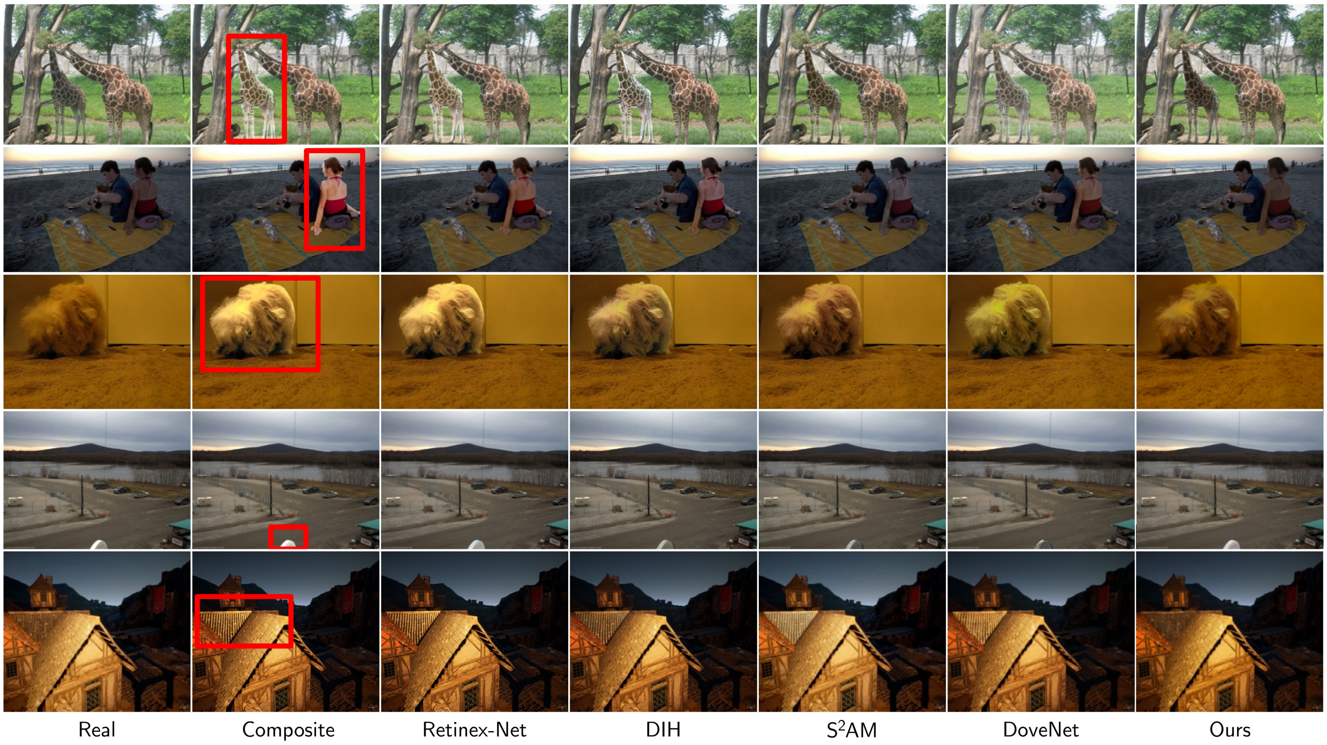

Figure 4. Qualitative comparison across four sub-datasets of iHarmony4 [11] and our new HVIDIT dataset (one example for each dataset). From top to bottom: HCOCO, HAdobe5k, HFlickr, Hday2night, and HVIDIT. Red boxes in composite images mark foreground.

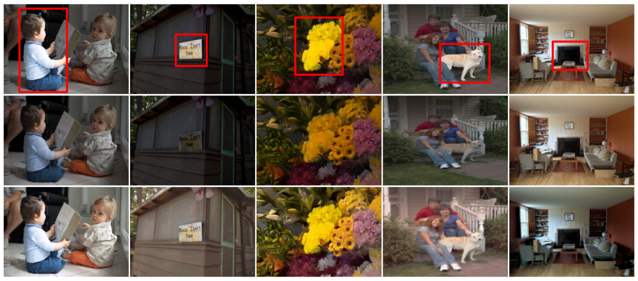

Figure 5. Image harmonization visual results with normal masks (middle row) and inverted masks (bottom row) on composite images (top row). Red boxes mark foreground of normal masks.

Method	PSNR↑	MSE↓	fMSE↓	SSIM↑	fSSIM↑
base	34.90	51.14	518.83	0.9599	0.8133
base+AdaIN	35.34	48.96	467.19	0.9627	0.8415
base+lighting(with l^{bg} only)	35.32	44.18	459.31	0.9604	0.8159
base+lighting	**35.82**	**40.62**	**428.21**	**0.9638**	**0.8447**

Table 2. Quantitative comparison about efficacy of light learning and transferring (lighting) on iHarmony4 dataset.

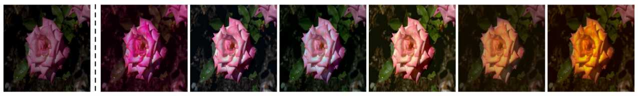

Figure 6. Changing light latent code of an image (left) from light learner produces different results in different lighting conditions.

Efficacy of Light Learning. We walk in the light latent space to see if our light learner has learned relevant light representation. Give an image, we use the light learner of our model (base+lighting) to extract light latent code of this input image as l_{fg} (consider this whole image as foreground), and change this code to obtain l_{bg}, then produce the "harmonized" result by recovery as output. Figure 6 illustrates an example with outputs under different lighting conditions, indicating the efficacy of our light learner.

Efficacy of Light Transferring. We design an experiment to employ our model for transferring the light from one source image to another target image, as shown in Figure 7, and the results show that source light is successfully transferred to target image thanks to our light transferer.

4.4. Analysis of Guiding

We further conduct ablation study to validate the efficacy of our inharmony-free patch relation modeling (guiding) as follows: (1) our method without lighting and guiding (base), (2) guiding reflectance and illumination by computing patch relations on composite images using SSIM (composite), (3) our guiding reflectance only (reflectance), (4) our guiding illumination only (illumination), (5) our inharmony-free guiding, and (6) guiding by computing patch relations on real images using SSIM as ground-truth

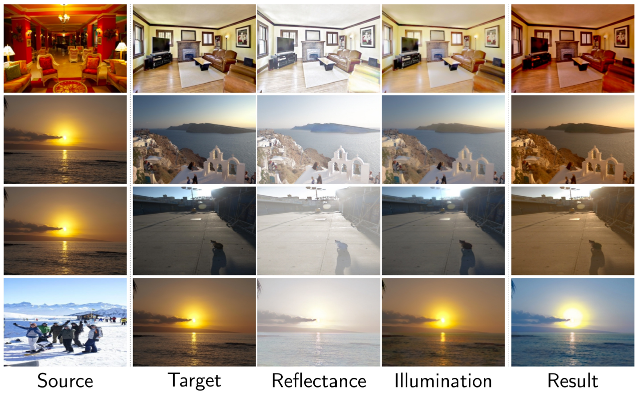

Figure 7. Our model can transfer the light from source to target. Reflectance and illumination are disentangled from target.

Method	PSNR↑	MSE↓	fMSE↓	SSIM↑	fSSIM↑
base	34.90	51.14	518.83	**0.9599**	0.8133
base+guiding(composite)	33.87	73.93	692.58	0.9579	0.7986
base+guiding(reflectance)	34.76	49.26	492.32	0.9581	0.8160
base+guiding(illumination)	34.16	49.61	504.72	0.9570	0.8033
base+guiding	**34.96**	**47.07**	**484.62**	0.9598	**0.8208**
base+guiding(real)	35.88	38.51	397.21	0.9616	0.8286

Table 3. Quantitative comparison about efficacy of inharmony-free patch relation modeling (guiding) on iHarmony4 dataset.

for comparison. The results listed in Table 3 demonstrate that, guiding by composite images misleads the intrinsic image harmonization, our guiding on reflectance or illumination improves performance, while our guiding boosts performance significantly, indicating the efficacy of our inharmony-free learning. Even so, taking real image guiding as a reference, there is still room for improvement in the guiding way, which we leave for future work.

4.5. Experiment on Our New HVIDIT Dataset

Following [49, 11], we merge the training set of our HVIDIT into iHarmony4 to retrain the models, and evaluate the models on both test sets, yielding results in Table 4, which draws the same conclusion as Table 1. For inharmonious triangle roof (foreground) in composite image of the fifth row in Figure 4, previous methods are hard to adjust foreground light to be close to background light, while our method performs the best, with similar light as the real image. We include more visual results in *supplementary file*.

4.6. Experiment on Real Composite Images

Following [49, 12, 11], we conduct user study on 99 real composite images provided by [49] for subjective evaluation. In result, we invite 60 subjects to participate in user study and acquire a total of 29700 pairwise results for all 99 images, with 30 results for each pair of different methods on average. All subjects are not aware of this image harmonization task, and are only required to select the visually better one corresponding to better method for each pair, then, we record how many times one method is selected in each pair on all 99 images as the statistics for pairwise comparison of Bradley-Terry (B-T) model [7, 29], to calculate global ranking score for each method. Table 5 and Figure 8 report B-T score and visual comparison respectively with three state-of-the-art image harmonization methods, and our method still achieves best performance with highest B-T score and best visual effect. Please refer to *supplementary file* for visual comparison of different methods to harmonize 99 real composite images.

Dataset	HVIDIT			All		
	PSNR↑	MSE↓	fMSE↓	PSNR↑	MSE↓	fMSE↓
Composite	38.53	53.12	1604.41	31.92	167.39	1386.12
Retinex-Net [52]	36.32	53.01	1603.21	31.08	165.09	1381.32
DIH [49]	36.62	45.55	1207.03	32.65	80.37	800.73
S^{2}AM [12]	36.24	45.82	1230.92	33.86	53.88	594.90
DoveNet [11]	36.80	35.36	1186.19	34.68	51.88	541.74
Ours (base)	40.55	33.16	934.63	35.09	46.76	512.05
Ours (base+lighting)	40.31	22.51	861.09	35.97	37.17	411.74
Ours (base+guiding)	40.29	25.57	925.01	35.78	42.48	470.30
Ours	**41.55**	**20.16**	**800.92**	**35.99**	**35.61**	**390.03**

Table 4. Quantitative comparison on our new HVIDIT dataset. We report results of iHarmony4 dataset in *supplementary file*.

Method	Composite	DIH [49]	S^{2}AM [12]	DoveNet [11]	Ours
B-T score↑	0.582	0.884	1.026	1.146	**1.735**

Table 5. User study comparison on 99 real composite images.

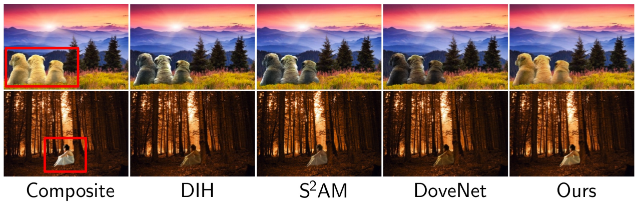

Figure 8. Visual comparison to harmonize real composite images.

5. Conclusion

In this paper, we propose a novel way of harmonizing composite images, namely, intrinsic image harmonization, aiming to eliminate the inharmony via separable reflectance and illumination intrinsic image harmonization. We respectively devise lighting and guiding to transfer the light from background to foreground and learn inharmony-free patch relations for better reflectance and illumination harmonization. Both extensive experiments and ablation studies demonstrate the power of our method and the efficacy of each component. Besides, we also contribute a new challenging harmonization dataset for specifically benchmarking illumination harmonization. We hope that our work opens up new avenues for image harmonization.

References

[1] Kenneth R Alexander and Michael S Shansky. Influence of hue, value, and chroma on the perceived heaviness of colors. *Perception & Psychophysics*, 19:72–74, 1976. 4

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM TOG*, 28(3):24, 2009. 1, 5

[3] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE TPAMI*, 37(8):1670–1687, 2014. 1, 2, 4

[4] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2(3-26):2, 1978. 1, 2, 3

[5] Anil S Baslamisli, Hoang-An Le, and Theo Gevers. Cnn based learning using reflection and retinex models for intrinsic image decomposition. In *CVPR*, pages 6674–6683, 2018. 6

[6] Sai Bi, Kalyan Sunkavalli, Federico Perazzi, Eli Shechtman, Vladimir G Kim, and Ravi Ramamoorthi. Deep cg2real: Synthetic-to-real translation via image disentanglement. In *ICCV*, pages 2730–2739, 2019. 2

[7] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 8

[8] David H Brainard and Brian A Wandell. Analysis of the retinex theory of color vision. *Journal of the Optical Society of America A*, 3(10):1651–1661, 1986. 2

[9] Matthew Brown and David G Lowe. Automatic panoramic image stitching using invariant features. *IJCV*, 74(1):59–73, 2007. 1

[10] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. BargainNet: Background-guided domain translation for image harmonization. *arXiv preprint arXiv:2009.09169*, 2020. 2

[11] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. DoveNet: Deep image harmonization via domain verification. In *CVPR*, pages 8394–8403, 2020. 1, 2, 5, 6, 7, 8

[12] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE TIP*, 29:4759–4771, 2020. 1, 2, 5, 6, 8

[13] Majed El Helou, Ruofan Zhou, Johan Barthas, and Sabine Süsstrunk. VIDIT: Virtual image dataset for illumination transfer. *arXiv preprint arXiv:2005.05460*, 2020. 5

[14] Majed El Helou, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, et al. AIM 2020: Scene relighting and illumination estimation challenge. In *ECCVW*, 2020. 5

[15] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, pages 8944–8952, 2018. 2, 3, 6

[16] Xueyang Fu, Yinghao Liao, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A probabilistic method for image enhancement with simultaneous illumination and reflectance estimation. *IEEE TIP*, 24(12):4965–4977, 2015. 2

[17] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *CVPR*, pages 2782–2790, 2016. 2

[18] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *ICCV*, pages 349–356, 2009. 2

[19] Hossein Bakhshi Golestani and Mohammad Ghanbari. Window size influence on ssim fidelity. In *7'th International Symposium on Telecommunications (IST'2014)*, pages 355–360. IEEE, 2014. 5

[20] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, pages 2335–2342, 2009. 2, 3, 4

[21] Xiaojie Guo, Yu Li, and Haibin Ling. LIME: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 26(2):982–993, 2016. 2

[22] Kaiming He and Jian Sun. Statistics of patch offsets for image completion. In *ECCV*, pages 16–29, 2012. 2

[23] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 6

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

[25] Micah K Johnson, Kevin Dale, Shai Avidan, Hanspeter Pfister, William T Freeman, and Wojciech Matusik. Cg2real: Improving the realism of computer generated images using a large collection of photographs. *IEEE TVCG*, 17(9):1273–1285, 2010. 2

[26] Ron Kimmel, Michael Elad, Doron Shaked, Renato Keshet, and Irwin Sobel. A variational framework for retinex. *IJCV*, 52(1):7–23, 2003. 2, 3, 4

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[28] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM TOG*, 33(4):1–11, 2014. 5

[29] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *CVPR*, pages 1701–1709, 2016. 8

[30] Jean-Francois Lalonde and Alexei A Efros. Using color compatibility for assessing image realism. In *ICCV*, pages 1–8, 2007. 1, 2

[31] Edwin H Land. The retinex theory of color vision. *Scientific American*, 237(6):108–129, 1977. 2

[32] Edwin H Land. An alternative technique for the computation of the designator in the retinex theory of color vision. *Proceedings of the National Academy of Sciences*, 83(10):3078–3080, 1986. 2

[33] Edwin H Land and John J McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61(1):1–11, 1971. 2, 3

[34] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for ir-

regular holes using partial convolutions. In *ECCV*, pages 85–100, 2018. 1

[35] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. In *CVPR*, pages 3248–3257, 2020. 2, 3, 6

[36] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *CVPR*, pages 5124–5133, 2020. 2

[37] Michael T Orchard, Charles A Bouman, et al. Color quantization of images. *IEEE Transactions on Signal Processing*, 39(12):2677–2690, 1991. 4

[38] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH*, pages 313–318, 2003. 1, 2

[39] Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A Efros, and George Drettakis. Multi-view relighting using a geometry-aware network. *ACM TOG*, 38(4):1–14, 2019. 2

[40] François Pitié, Anil C Kokaram, and Rozenn Dahyot. N-dimensional probability density function transfer and its application to color transfer. In *ICCV*, pages 1434–1439, 2005. 2

[41] Min-cheol Sagong, Yong-goo Shin, Seung-wook Kim, Seung Park, and Sung-jea Ko. Pepsi: Fast image inpainting with parallel decoding network. In *CVPR*, pages 11360–11368, 2019. 1, 5

[42] Robert J Schalkoff. *Digital Image Processing and Computer Vision*, volume 286. Wiley New York, 1989. 4

[43] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. SfSNet: Learning shape, reflectance and illuminance of faces 'in the wild'. In *CVPR*, pages 6296–6305, 2018. 2

[44] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *CVPR*, pages 9243–9252, 2020. 1

[45] Wu Shi, Chen Change Loy, and Xiaoou Tang. Deep specialized network for illuminant estimation. In *ECCV*, pages 371–387, 2016. 2

[46] Louise L Sloan. Vision: (Value, chroma, and hue). *Psychological Bulletin*, 24(2):100, 1927. 4

[47] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM TOG*, 29:1–10, 2010. 1, 2

[48] Michael W Tao, Micah K Johnson, and Sylvain Paris. Error-tolerant image compositing. *IJCV*, 103(2):178–189, 2013. 2

[49] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, pages 3789–3797, 2017. 1, 2, 5, 6, 8

[50] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, pages 6849–6857, 2019. 2

[51] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 5

[52] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018. 2, 6, 8

[53] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3D objects from images in the wild. In *CVPR*, pages 1–10, 2020. 2

[54] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM TOG*, 31(4):1–10, 2012. 1, 2

[55] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, pages 5505–5514, 2018. 1, 5

[56] Julio Zaragoza, Tat-Jun Chin, Michael S Brown, and David Suter. As-projective-as-possible image stitching with Moving DLT. In *CVPR*, pages 2339–2346, 2013. 1

[57] Fan Zhang and Feng Liu. Parallax-tolerant image stitching. In *CVPR*, pages 3262–3269, 2014. 1

[58] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE TPAMI*, 21(8):690–706, 1999. 1

[59] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *ICCV*, pages 7194–7202, 2019. 2

[60] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. Learning a discriminative model for the perception of realism in composite images. In *CVPR*, pages 3943–3951, 2015. 1, 2

[61] Maria Zontak and Michal Irani. Internal statistics of a single natural image. In *CVPR*, pages 977–984, 2011. 2, 5