

Intrinsic Harmonization for Illumination-Aware Compositing

Chris Careaga
Simon Fraser University

S. Mahdi H. Miangoleh
Simon Fraser University

Yağız Aksoy
Simon Fraser University

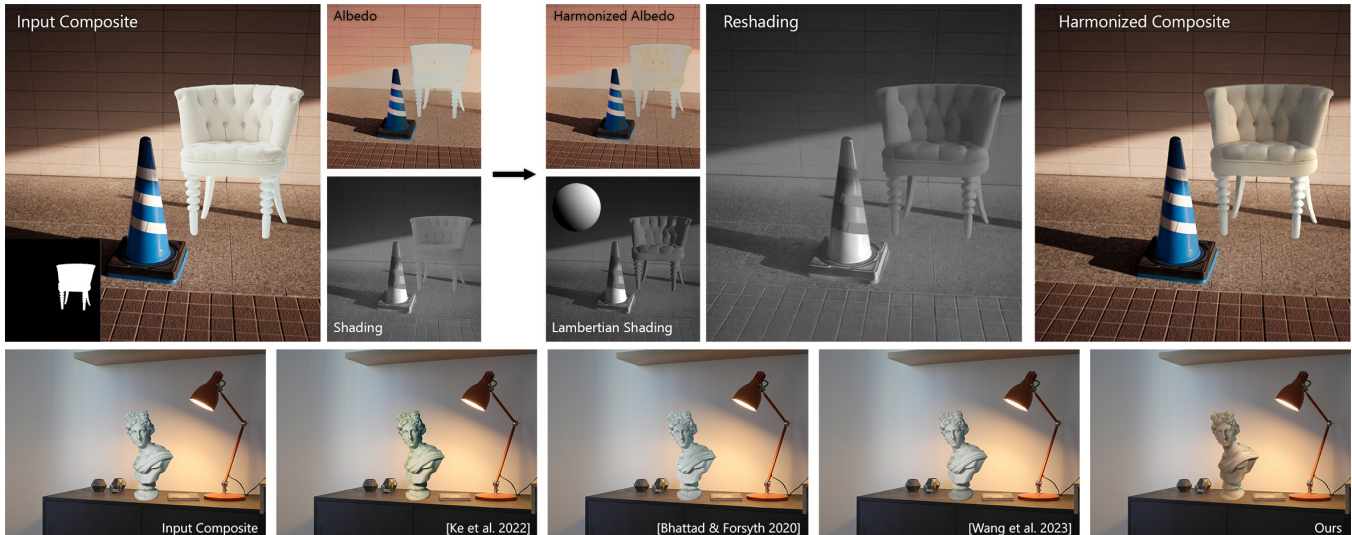


Figure 1: We propose an illumination-aware image harmonization approach for in-the-wild imagery. Our method is formulated in the intrinsic image domain. We use off-the-shelf networks to generate albedo, shading and surface normals for the input composite and background image. We first harmonize the albedo of the background and foreground by predicting image editing parameters. Using normals and shading we estimate a simple lighting model for the background illumination. With this lighting model, we render Lambertian shading for the foreground and refine it using a network trained on segmentation datasets via self-supervision. When compared to prior works we are the only method that is capable of modeling realistic lighting effects. Image credit: Unsplash users mak_jp, Daniil Silantev, Jean-Philippe Delberghe and Luo Dan

ABSTRACT

Despite significant advancements in network-based image harmonization techniques, there still exists a domain disparity between typical training pairs and real-world composites encountered during inference. Most existing methods are trained to reverse global edits made on segmented image regions, which fail to accurately capture the lighting inconsistencies between the foreground and background found in composited images. In this work, we introduce a self-supervised illumination harmonization approach formulated in the intrinsic image domain. First, we estimate a simple global lighting model from mid-level vision representations to generate a rough shading for the foreground region. A network then refines this inferred shading to generate a harmonious re-shading that

aligns with the background scene. In order to match the color appearance of the foreground and background, we utilize ideas from prior harmonization approaches to perform parameterized image edits in the albedo domain. To validate the effectiveness of our approach, we present results from challenging real-world composites and conduct a user study to objectively measure the enhanced realism achieved compared to state-of-the-art harmonization methods.

CCS CONCEPTS

• **Computing methodologies** → **Image manipulation.**

KEYWORDS

image compositing, object relighting, intrinsic decomposition, self-supervised learning, image harmonization

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SA Conference Papers '23, December 12–15, 2023, Sydney, NSW, Australia
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0315-7/23/12...\$15.00
<https://doi.org/10.1145/3610548.3618178>

ACM Reference Format:

Chris Careaga, S. Mahdi H. Miangoleh, and Yağız Aksoy. 2023. Intrinsic Harmonization for Illumination-Aware Compositing. In *SIGGRAPH Asia 2023 Conference Papers (SA Conference Papers '23)*, December 12–15, 2023, Sydney, NSW, Australia. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3610548.3618178>

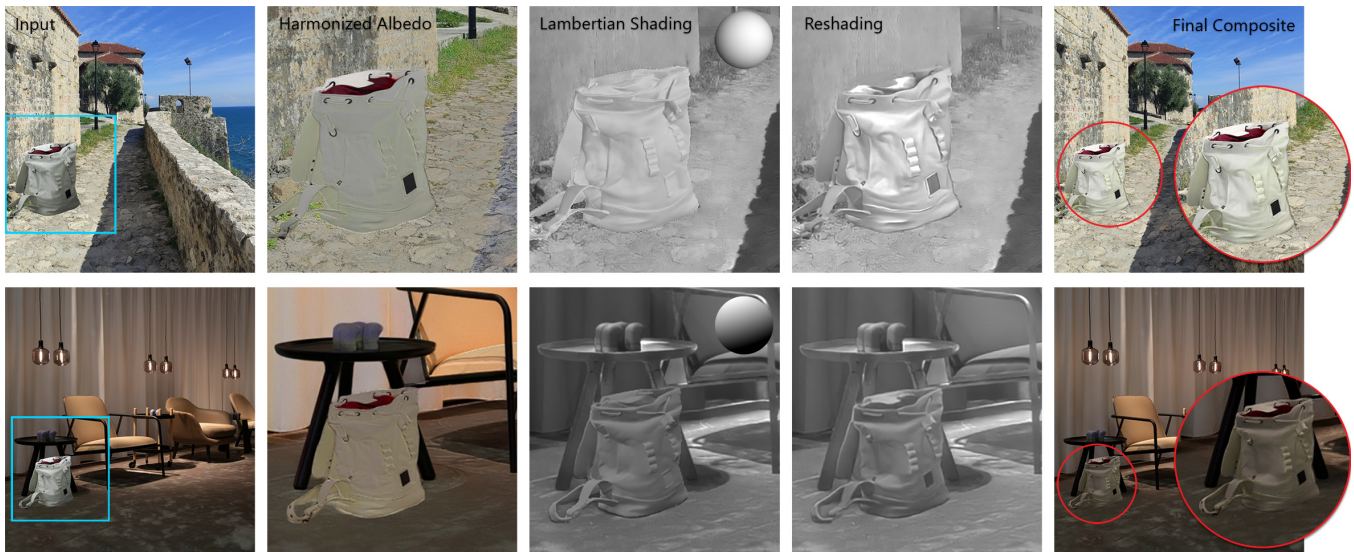


Figure 2: Our method is able to relight foreground regions under varying background lighting conditions. We estimate a lighting model of the background to render a novel shading that is refined by our re-shading network. By harmonizing the albedo, we are also able to represent illumination color. Image credit: Unsplash users Yuliya Yevseyeva, Jean-Philippe Delberghe and _k8_

1 INTRODUCTION

Compositing, or inserting an object onto a novel background, is an important image editing task that requires the object to naturally blend in the new environment. This requires the appearance of the inserted object to be readjusted to fit the background in a process called image harmonization. For a realistic composite, the harmonized object should match the color content as well as the illumination present in the background.

Image harmonization has been widely studied in the literature as an image editing problem where self-supervised methods are used to estimate a set of color and tone adjustment operations to match the color contents of the object with the background. Relighting the object to match the illumination in the novel environment, however, has been largely neglected in the harmonization literature. This comes from the difficulty of realistically relighting an object in-the-wild, which requires an accurate estimation of the illuminating environment as well as a detailed and accurate geometry. As a result, most relighting methods in the literature focus on a specific domain such as portraits [Pandey et al. 2021; Yeh et al. 2022]. Restricting the domain of relighting makes it possible to generate training data in controlled environments for image-to-image relighting. However, this makes relighting methods in the literature inapplicable to the general problem of image compositing.

In this work, we model the image harmonization problem in the intrinsic domain. Intrinsic image decomposition is a fundamental mid-level vision problem that represents an image as the product of the reflectance of the materials and the effect of illumination in the scene:

$$I = S \cdot A, \quad (1)$$

where S and A represent the shading and albedo, respectively. By isolating the scene colors contained in the albedo from shading,

the intrinsic representation allows us to divide the image harmonization problem into two: color harmonization and relighting. In our harmonization pipeline, we first harmonize the color content of the foreground to match the background in the albedo space by borrowing ideas from the rich color harmonization literature. We then turn our focus onto the challenging relighting problem in the shading domain.

Our aim is to generate a new shading for the composited object that reflects the new illumination environment. For this purpose, we generate an initial estimation of the shading using a simple Lambertian shading model and surface normals estimated for the background and the inserted object. We first estimate the illumination environment of the background using a simplified parametric illumination model. We then generate the Lambertian shading for the inserted object and composite this shading map onto the original shading of the background, representing a starting point for our re-shading network. Together with the RGB composite, we use the initial shading as input and train our network to generate a realistic new shading for the object. This allows us to train our re-shading network in a self-supervised manner using segmentation datasets.

We show that by dividing the harmonization problem into two and formulating a self-supervised relighting method, realistic composite images can be generated where the inserted object not only reflects the color content of the image but also matches the illumination present in the background. As shown in Figures 1 and 2, this allows us to generate much more realistic composite images when compared to prior work on image harmonization. Through qualitative examples and a subjective evaluation, we demonstrate the performance of our method in challenging scenarios.

2 RELATED WORK

Image harmonization. Previously proposed image harmonization methods have predominantly been trained in a self-supervised manner in order to undo various image edits performed on a specific region of a natural image [Cong et al. 2022, 2020; Ke et al. 2022; Xue et al. 2022]. These edits only represent image-level differences (brightness, saturation, hue, etc.) between the harmonized ground truth and the unharmonized input. This creates a gap between training and inference as real-world composites require more complex operations to be properly harmonized. Various approaches have been proposed to model the image harmonization problem with more realistic assumptions, including illumination harmonization. Liao et al. [2019] propose a perceptually-inspired shading model to relight image segments. Their approach first estimates a simple geometric model that can be reshaded. They then utilize low-level algorithms to estimate shading effects that are not represented by the rough geometry. Similarly, our approach proposes to utilize estimated geometry that can be reshaded, but replaces the approximate shading model with a data-driven shading refinement process that can be trained via self-supervision. The method of Guo et al. [2022, 2021] formulates an approach based on the intrinsic image domain to jointly separate and harmonize albedo and shading. Their method often fails to generate novel shading due to a lack of proper training data and learning to harmonize and perform intrinsic decomposition with one network. Hu et al. [2021] develop a generated dataset and method to relight humans in outdoor scenes. Their method focuses on a specific use case and requires ground-truth geometry as input, making it difficult to use in the wild. Similarly, Bao et al. [2022] propose a dataset and method to perform illumination harmonization, but they mainly focus on relighting objects on a flat ground plane in outdoor scenes. The work of Bhattad and Forsyth [2022] formulates an intrinsic approach that re-shades a masked object using Deep Image Prior. Their method fails to generate realistic shading estimations and requires multiple minutes of run time for a single image. Wang et al. [2023] propose a simpler method of modeling illumination changes in image harmonization, but their method can only locally modulate color-based editing of the foreground region. Given this limitation and lack of explicit reasoning about albedo, they are not able to fully harmonize composites with major differences in lighting between foreground and background.

Our method, on the other hand, borrows ideas from typical image harmonization approaches while also fully modeling the illumination mismatch present in real-world composite images. Using our self-supervised training approach that allows us to employ large-scale segmentation datasets, we are able to perform realistic relighting of foreground regions that match the background lighting environment for in-the-wild composites.

Object insertion. Rather than attempting to composite image segments into novel scenes, the task of object insertion attempts to insert a 3D object into a 2D photograph of a scene. This is typically accomplished by *inverse rendering* 2D scenes to infer characteristics of the 3D scene they depict. Lopez-Moreno et al. [2010] proposed an algorithm to recover lighting information from an image segment. They use their lighting information to relight recovered 3D geometry of an object making it appear as if it belongs in the scene.

Later works take this idea further, recovering a 3D representation of a scene complete with lighting that can then be used to re-render a 3D object using a typical rendering pipeline [Karsch et al. 2011, 2014]. With the advent of deep learning, inverse rendering techniques have become data-driven, leveraging large-scale synthetic datasets of indoor scenes to train networks for estimating scene intrinsics [Li et al. 2020, 2022]. Our method, on the other hand, does not assume a 3D object as input and instead works completely in the 2D image domain. Additionally, our method is able to work both indoors and outdoors while not requiring large-scale datasets for inferring scene characteristics and instead makes use of off-the-shelf networks for mid-level vision tasks.

Image relighting. Given the complexity of image relighting, prior methods mainly focus on a specific use case such as portraits [Pandey et al. 2021; Yeh et al. 2022] or outdoor structures [Griffiths et al. 2022]. These methods rely on large-scale, difficult-to-obtain datasets, or multi-view scenes [Nicolet et al. 2020; Philip et al. 2019, 2021], in order to achieve realistic results.

Additionally, if the target illumination is not provided, as is the case in image harmonization, lighting must be predicted. Generating estimations of lighting configuration is a difficult task also relying on hard-to-capture datasets [Garon et al. 2019; Zhang et al. 2019]. By leveraging off-the-shelf mid-level estimations, our method learns to relight the foreground from a single image, in the wild, without requiring ground-truth source and target illumination examples. Additionally, we utilize a simple lighting estimation formulation to guide our re-shading network, while still generating realistic shading estimations.

3 INTRINSIC HARMONIZATION

Our approach makes use of multiple off-the-shelf methods to generate mid-level representations. We denote the background scene and the foreground object to be composited as I_b and I_f , respectively. We start by generating the intrinsic decomposition of both images using the method by Careaga and Aksoy [2023]:

$$I^f = S^f \cdot A^f, \quad I^b = S^b \cdot A^b. \quad (2)$$

Where S and A denote the single channel shading and the RGB albedo, respectively. We define the compositing in the albedo and the shading separately:

$$A^c = \alpha A^f + (1 - \alpha) A^b, \quad S^c = \alpha S^f + (1 - \alpha) S^b, \quad I^c = A^c \cdot S^c. \quad (3)$$

where α represents the foreground mask, and I^c , A^c , and S^c represent the composite image and its intrinsic components. Our approach uses linear RGB when performing any albedo and shading operations as Equation 2 assumes linear RGB values. When given a standard RGB image as input we reverse the gamma-correction process using a gamma value of 2.2. This separation of the compositing problem allows us to perform color and illumination harmonization in two separate steps. We will first give an overview of our albedo harmonization, and then detail our self-supervised approach to generate the new foreground shading in Section 4. Examples of our harmonized albedo can be seen in Figures 2 and 3.

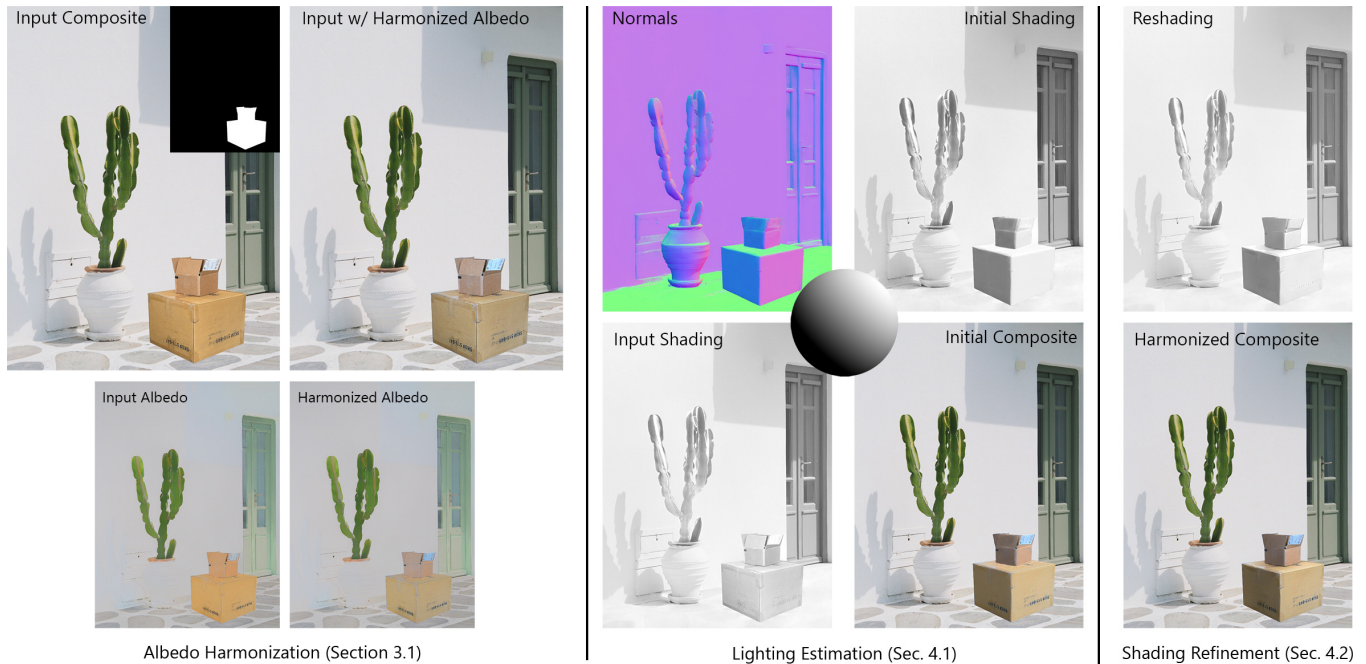


Figure 3: Our pipeline consists of three main parts. We first perform harmonization on the estimated albedo of the scene (Section 3.1). We next use estimated normals and shading of the background to optimize the parameters of a simple lighting model (Section 4.1). Finally, we use the lighting model to render a Lambertian shading of the foreground region. We refine this Lambertian shading with a network to generate a final realistic shading layer that we combine with our harmonized albedo to create our final result (Section 4.2). Image credit: Unsplash user Chloé Chavanon

3.1 Albedo Harmonization

The color content of a scene is represented in the albedo, therefore we design a simple color-based harmonization method to adjust the foreground albedo to harmonize well with the colors in the background albedo. Similar to prior parameter-based harmonization approaches [Ke et al. 2022; Wang et al. 2023], we aim to find a set of editing parameters that control common image editing operations such as changing the exposure, saturation, color curve, and white balance. We achieve this through a self-supervised setup using segmentation datasets. For a given segment, we apply a random set of operations to the object to create a mismatch between the simulated *composited* albedo and the original image. We then train a network to estimate the editing parameters that will re-create the original appearance of the object in the scene, A^c . For this purpose, we utilize the editing network proposed by Miangoleh et al. [2023]. We train our network with the mean squared error defined on the RGB albedo using the MS COCO [Lin et al. 2014] and Davis [Perazzi et al. 2016] datasets for 100 epochs. We provide a detailed description of our setup in the supplementary material.

4 FOREGROUND RE-SHADING

Harmonizing the illumination of the foreground region in a composite image can be posed as a specific case of image relighting. Achieving physically-accurate relighting would require an accurate HDRI illumination environment together with a highly detailed geometry to physically re-render the object. However, the creation

of a relighting dataset for data-driven relighting is highly challenging and hard to scale due to the controlled setup such a dataset requires. Instead of training a network for image-to-image relighting, in order to create a novel foreground shading that matches the background environment, we develop a simple Lambertian shading model with a parametric illumination representation. We then train a network to refine our Lambertian shading in order to generate a realistic shading map for the final composite image. We show that by defining the relighting problem as the refinement of the Lambertian shading, we can train our network in a self-supervised manner using standard segmentation datasets.

4.1 Parametric Illumination Model

We first start our re-shading pipeline by estimating a simple illumination model for the background scene using the Lambertian shading model. We define our illumination model as a combination of a directional light source represented by the vector $\vec{l} \in \mathbb{R}^3$ and a constant ambient illumination c . The Lambertian shading model represents the shading of a scene using the surface normals and the illumination environment. Using our simple illumination model, the Lambertian shading for the background \tilde{S}^b is defined as:

$$\tilde{S}_i^b = \vec{n}_i^b \cdot \vec{l} + c, \quad (4)$$

where \vec{n}_i^b represents the surface normal at pixel i estimated using an off-the-shelf method [Eftekhari et al. 2021]. We estimate the

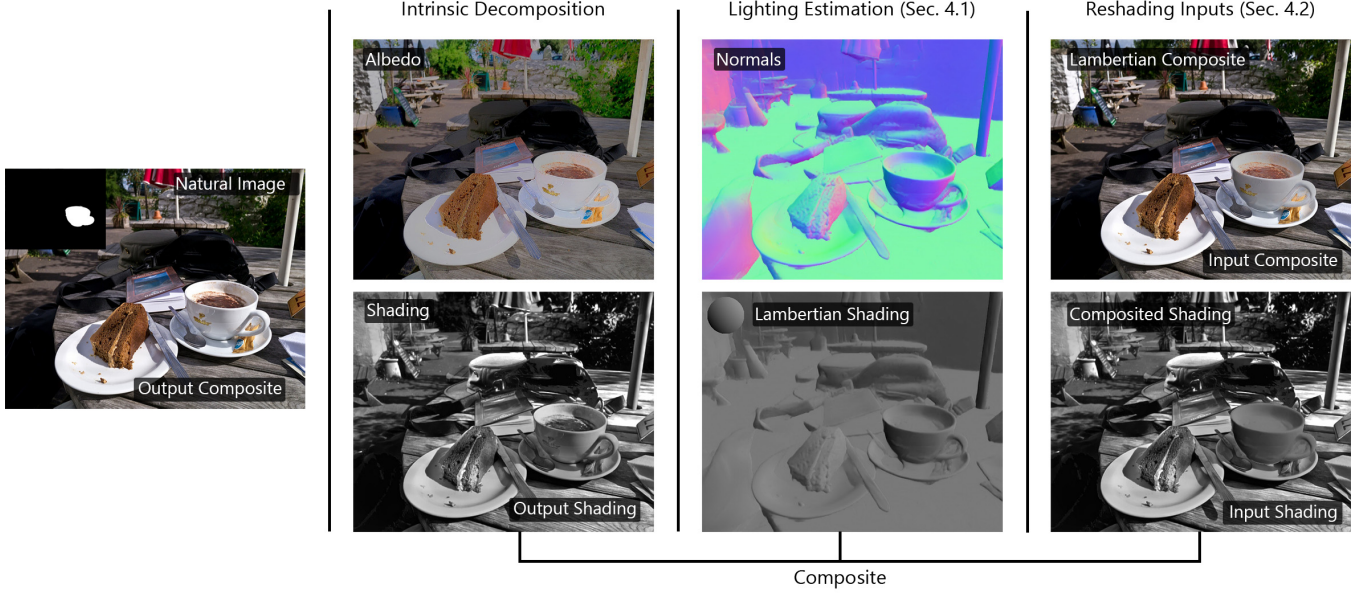


Figure 4: To train our re-shading network (Section 4.2) we propose a self-supervised data generation strategy. We use image segmentation datasets to provide masked image regions for diverse scenes. We first perform intrinsic decomposition on the image. We use our lighting estimation method (Section 4.1) to generate lighting parameters for the non-masked region. We render Lambertian for the foreground region and composite onto the background shading. This shading is used, along with the original albedo, to generate a new composite with Lambertian shading in the masked region. Our model is trained to reverse this process, learning to map the Lambertian shading to accurate and realistic shading. Image credit: Flickr user shirokazan

parameters of our lighting model with a least squares optimization:

$$(\vec{l}, c) = \arg \min_{\vec{v}, x} \sum_i (S_i^b - \vec{n}_i^b \cdot \vec{v} + x)^2. \quad (5)$$

In other words, we search for the setting of \vec{l} and c such that the rendered Lambertian shading best reconstructs the estimated shading of the background scene. In order to ensure the lighting model parameters take on plausible values, we constrain the values of \vec{l} and c to be positive. This keeps \vec{l} in the outward-facing hemisphere. We solve this minimization problem using gradient-based optimization with the Adam optimizer [Kingma and Ba 2015]. We show an example of our lighting estimation pipeline in Figure 5.

We use the estimated illumination model to generate the Lambertian shading for the foreground object:

$$\tilde{S}_i^f = \vec{n}_i^f \cdot \vec{l} + c, \quad (6)$$

which represents the starting point for our shading refinement process. We composite \tilde{S}^f onto S^b to generate the initial composite shading \tilde{S}^c , which we use as input for our shading refinement.

4.2 Shading Refinement

We define the re-shading problem as the refinement of \tilde{S}^c into a realistic shading map S^c such that when multiplied with the albedo, reconstructs the illumination-harmonized composite image. This definition allows us to train our re-shading network in a self-supervised fashion. We generate our input - ground-truth pairs using real images with a segmentation mask, representing the

composited region. We then estimate the Lambertian shading of the object using our parametric model, which is used as the input to our network. The ground-truth shading is then defined as the shading estimated for the original image S^c . The input to our network is the foreground mask, the RGB composite image with Lambertian shading, $A^c \cdot \tilde{S}^c$, the Lambertian shading \tilde{S}^c , as well as the surface normals and a depth map generated using methods by Eftekhari et al. [2021] and Miangoleh et al. [2021] to provide our network with geometric context. We channel-wise concatenate the inputs for a $h \times w \times 9$ input, h and w representing the height and width of the image, respectively. An overview of the self-supervised data generation process can be seen in Figure 4. We supervise our network using losses defined on the shading and the final reconstructed RGB image, $I^c = A^c \cdot S^c$. We use the mean squared error:

$$\mathcal{L}_s = \text{MSE}(\hat{S}^c, S^c), \quad \mathcal{L}_i = \text{MSE}(\hat{I}^c, I^c), \quad (7)$$

where \hat{S}^c is the estimated shading and $\hat{I}^c = A^c \cdot \hat{S}^c$ is the estimated composite image. We also use the multi-scale gradient loss [Li and Snavely 2018] as an edge-aware smoothness loss:

$$\mathcal{L}_{sg} = \sum_m \text{MSE}(\nabla \hat{S}^{c,m}, \nabla S^{c,m}), \quad \mathcal{L}_{ig} = \sum_m \text{MSE}(\nabla \hat{I}^{c,m}, \nabla I^{c,m}), \quad (8)$$

where $\nabla S^{c,m}$ represents the gradient of S^c at scale m . We combine these losses to compute our final loss used during training:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_i + \mathcal{L}_{sg} + \mathcal{L}_{ig} \quad (9)$$

using unit weights.

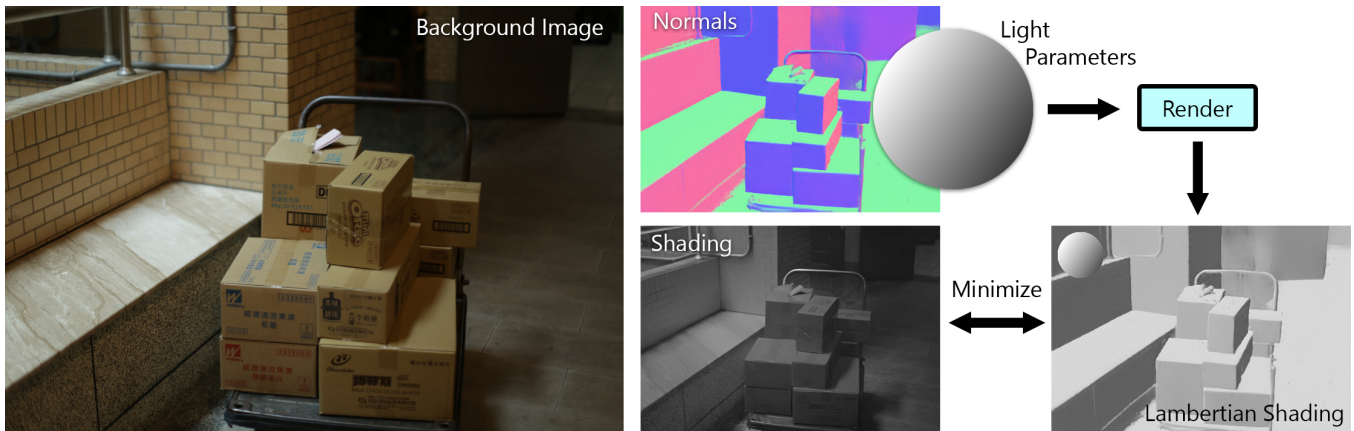


Figure 5: Using estimated normals and shading, we optimize lighting parameters such that the rendered Lambertian shading best matches the estimated shading of the scene. We use this lighting representation to infer the general shading of an object in the scene. Image credit: Unsplash user ej1209_

4.3 Network Architecture and Training

We combine 3 datasets to train our re-shading network: the COCO Dataset [Lin et al. 2014], a 50,000 image subset of the SA-1B Dataset [Kirillov et al. 2023], and the Multi-Illumination Dataset (MID) [Murrman et al. 2019]. For COCO and SA-1B, we utilize the provided segmentation masks and sample foreground segments that are sufficiently large. For MID, we use the segments provided in the dataset and use the provided multiple illuminations for dataset augmentation. We use the encoder-decoder architecture used by Ranftl et al. [2020] which consists of a ResNext101 [Xie et al. 2017] encoder and a RefineNet [Lin et al. 2017] decoder. We train our network using the Adam optimizer with a learning rate of 1×10^{-5} for 2 million iterations.

Our definition of the relighting problem as shading refinement allows us to use generic segmentation datasets for self-supervised training. As these datasets provide diverse training data for both indoor and outdoor conditions, we are able to perform illumination harmonization, i.e. object relighting in-context, in-the-wild.

5 EXPERIMENTAL EVALUATION

To evaluate the realism of our generated composites we compare against various state-of-the-art image harmonization methods. We focus our evaluation on methods that also attempt to model illumination harmonization. We perform a user study comparing multiple methods on difficult composite images. Additionally, we provide qualitative comparisons of the scenes used in our user study.

5.1 User Study

In order to concretely evaluate the effectiveness of our harmonization, we perform a qualitative user study. We model our user study after that of Wang et al. [2023] and perform a *two alternatives forced choice* survey. We compare our method to three state-of-the-art methods. The method of Bhattad and Forsyth [2022] also performs illumination harmonization in the intrinsic domain. Wang et al. [2023] utilize a gain map to modulate image edits allowing them to

Table 1: Bradley-Terry [Bradley and Terry 1952] ranking scores of naive composites and state-of-the-art methods compared to our models. While our model without reshading performs comparably to Wang et al. [2023]’s model, our full model with reshading achieves the highest score with a significant gap.

Method	B-T Score \uparrow
Naive Composite	0.0933
Bhattad and Forsyth [2022]	0.0893
Ke et al. [2022]	0.1727
Wang et al. [2023]	0.2078
Ours (w/o reshading)	0.1906
Ours (full)	0.2485

model non-global edits. The work of Ke et al. [2022] proposes a parametric harmonization approach for high-definition images. Each of these methods can produce results at high-resolution making them suitable for qualitative comparison against our method.

We generate 50 composites using free-to-use images from Unsplash. We aim to create difficult examples with a mismatch in color and illumination between the foreground and background regions. Examples of our composited images are shown in Figure 7. For each composite, our method is compared to the naive input composite, three prior works, and our method without shading harmonization. We generate image pairs consisting of our method and each of the comparison methods for each image, resulting in $5 \times 50 = 250$ pairs. Each participant is given context about image compositing and told to “determine which image has the foreground object better matching the background environment”. They are then shown a random set of 50 pairs without duplicate photographs. We crowd-sourced our survey via word of mouth on social media. We obtained responses from 70 subjects, resulting in 3500 total comparisons. We follow prior harmonization works [Cong et al. 2020; Wang et al. 2023] and analyze our responses using the Bradley-Terry model [Bradley and Terry 1952] to generate global ranking scores for each method. The

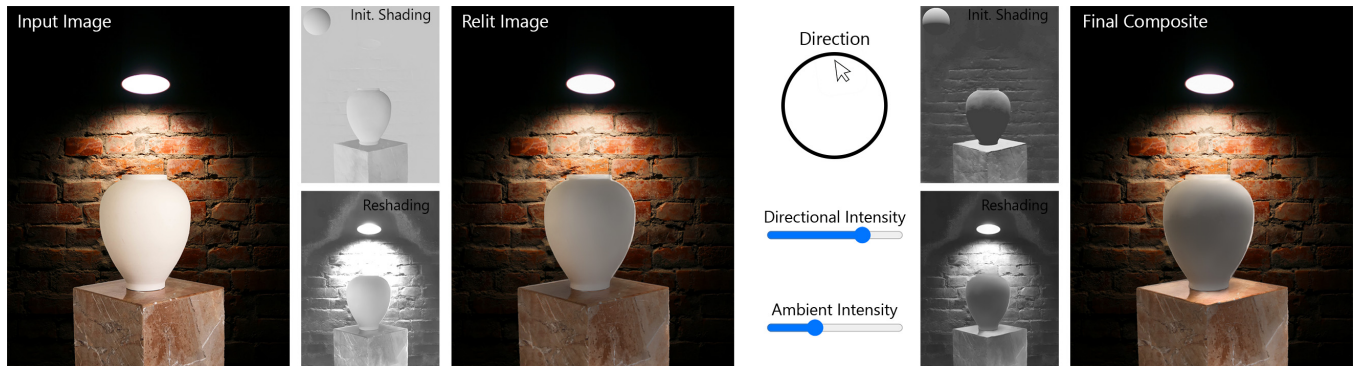


Figure 6: While our simple lighting model can fail when our assumptions are broken, due to its small set of interpretable parameters it can be easily edited by a human to achieve the desired result. This aspect of our approach could be incorporated into a GUI to interactively harmonize an image. Image credit: Unsplash users Mathilde Langevin and De an Sun

results of our user study are shown in Table 1. We observe that our method is preferred over all other methods. Additionally, we observe that our model is preferred significantly more when the illumination is harmonized, showing that illumination is an important aspect of realism when it comes to in-the-wild composites.

5.2 Qualitative Comparisons

Figure 7 shows qualitative comparisons on composite images gathered as part of our user study. We show that prior works fail to harmonize the illumination differences between the foreground and background. In the first row, none of the prior works are able to attenuate the shadow on the hat from the foreground’s original outdoor environment. In the second row, our method is able to estimate the bright outdoor light shining on the building facade in the background. In the third row, the prior works fail to soften the lighting on the box, leaving the original direct lighting that does not match the blue ambient lighting from the background. In the bottom row, our method is able to dim the overall brightness of the illumination on the boxes and also reflect the direction of the light coming from the window.

Figure 8 provides a comparison to prior works that also model intrinsics, or perform non-global edits to simulate altered lighting. We include the method of Guo et al. [2021] in our comparison as they also model the harmonization method in the intrinsic domain. We do not include their results in our user study as they can only estimate at a resolution of (256×256) . We observe that prior works struggle to estimate accurate intrinsic representations. The approach of Bhattad and Forsyth [2022] cannot generate high-frequency details in their shading due to estimating at low resolution, therefore their results typically don’t exhibit novel illumination effects. The work of Guo et al. [2021] attempts to perform intrinsic decomposition and harmonization jointly, but fails to estimate meaningful albedo and shading due to a lack of ground-truth supervision for these quantities. The method of Wang et al. [2023] predicts a gain map to modulate parametric edits. While their gain map does allow them to model local variations, their method cannot fully relight the foreground as they do not explicitly model albedo. This results in their harmonized foreground maintaining the illumination conditions from its source environment. Our method models

the problem with physical accuracy and can therefore predict a detailed and accurate re-shading of the foreground.

6 DISCUSSION

While we propose a pipeline for performing illumination-aware image harmonization, our re-shading network can also be utilized for interactive relighting. Given our simple lighting model, it is easy to alter the illumination conditions of the Lambertian shading given to our network. As Figure 2 shows, our re-shading network faithfully generates a realistic shading layer that obeys the conditions of the input Lambertian shading. This aspect of our method is also amenable to interactive image editing applications. As Figure 6 shows, in case our estimated parametric light model is inaccurate, a user can easily alter our lighting model by providing values for the lighting parameters. The user can then yield a more pleasing composite. When combined with our parametric albedo harmonization formulation, a user can have full control over the image harmonization process.

7 LIMITATIONS

Our method relies on multiple off-the-shelf networks to estimate various mid-level vision representations. Although our pipeline is generally robust to small errors in these estimations, an inaccurate intrinsic decomposition or surface normals may lead to inaccuracies in the lighting model estimation or in the Lambertian shading generated for the foreground. It does, however, mean that if these methods are improved, our model would see a similar gain in performance. Additionally, our lighting model is also rudimentary and cannot represent colorful illumination and hence we rely on the albedo harmonization network to account for any illumination color coming from the background environment. Given the assumptions of the Lambertian shading model, our lighting estimation can fail in difficult scenarios such as multiple colored light sources in the image. Finally, while we are able to relight the foreground object, our method does not model the cast shadows that the object may generate in the new environment. This requires a detailed understanding of the geometry of the background environment, and we believe this is a limitation to address in future work.

8 CONCLUSION

We introduce a self-supervised illumination harmonization approach formulated in the intrinsic image domain. We show that the intrinsic domain allows us to address two challenges in image harmonization, color and illumination mismatch, separately using dedicated models. We utilize ideas from the prior image harmonization literature to first match the color appearance of the foreground and background albedo by performing parameterized image edits. We then estimate a simple global lighting model from mid-level vision representations to generate Lambertian shading for the foreground region. We train a network via self-supervision to refine this inferred shading into a realistic re-shading that aligns with the background scene. This represents an interesting direction in the relighting literature, which typically models the problem as image-to-image translation trained on specialized datasets. To show the effectiveness of our approach, we conduct a user study on challenging real-world composites to objectively measure the enhanced realism achieved compared to state-of-the-art harmonization methods.

ACKNOWLEDGMENTS

The authors would like to thank Elise Saxon for helping with the creation of user study surveys and Ke Wang for promptly providing data and results for their method. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [RGPIN-2020-05375].

REFERENCES

- Zhongyun Bao, Chengjiang Long, Gang Fu, Daquan Liu, Yuanzhen Li, Jiaming Wu, and Chunxia Xiao. 2022. Deep Image-based Illumination Harmonization. In *Proc. CVPR*.
- Anand Bhattad and David A. Forsyth. 2022. Cut-and-Paste Object Insertion by Enabling Deep Image Prior for Reshading. *Proc. 3DV* (2022).
- Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39, 3/4 (1952), 324–345.
- Chris Careaga and Yağız Aksoy. 2023. Intrinsic Image Decomposition via Ordinal Shading. *SFU Tech. Rep.* (2023).
- Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. 2022. High-resolution image harmonization via collaborative dual transformations. In *Proc. CVPR*.
- Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. 2020. DoveNet: Deep Image Harmonization via Domain Verification. In *Proc. CVPR*.
- Ainaz Eftekhari, Alexander Sax, Jitendra Malik, and Amir Zamir. 2021. Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets From 3D Scans. In *Proc. ICCV*.
- Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. 2019. Fast Spatially-Varying Indoor Lighting Estimation. In *Proc. CVPR*.
- David Griffiths, Tobias Ritschel, and Julien Philip. 2022. OutCast: Single Image Relighting with Cast Shadows. *Comput. Graph. Forum* (2022).
- Zonghui Guo, Zhaorui Gu, Bing Zheng, Junyu Dong, and Haiyong Zheng. 2022. Transformer for Image Harmonization and Beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* (2022).
- Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. 2021. Intrinsic Image Harmonization. In *Proc. CVPR*.
- Zhongyun Hu, Ntumba Elie Nsambi, Xue Wang, and Qing Wang. 2021. NeurSF: Neural Shading Field for Image Harmonization. arXiv:2112.01314 [cs.CV]
- Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. 2011. Rendering Synthetic Objects into Legacy Photographs. *ACM Trans. Graph.* 30, 6 (2011).
- Kevin Karsch, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fong, Michael Sittig, and David Forsyth. 2014. Automatic Scene Inference for 3D Object Compositing. *ACM Trans. Graph.* 33, 3 (2014).
- Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson W.H. Lau. 2022. Harmonizer: Learning to Perform White-Box Image and Video Harmonization. In *Proc. ECCV*.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *Proc. ICLR*.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. arXiv:2304.02643 [cs.CV]
- Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2020. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In *Proc. CVPR*.
- Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. 2022. Physically-Based Editing of Indoor Scene Lighting from a Single Image. In *Proc. ECCV*.
- Zhengqi Li and Noah Snavely. 2018. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In *Proc. CVPR*.
- Zicheng Liao, Kevin Karsch, Hongyi Zhang, and David Forsyth. 2019. An Approximate Shading Model with Detail Decomposition for Object Relighting. *Int. J. Comput. Vision* 127 (2019).
- G. Lin, A. Milan, C. Shen, and I. Reid. 2017. RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. In *Proc. CVPR*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proc. ECCV*.
- Jorge Lopez-Moreno, Sunil Hadap, Erik Reinhard, and Diego Gutierrez. 2010. Compositing Images through Light Source Detection. *Computers & Graphics* 34, 6 (2010), 698–707.
- S. Mahdi H. Miangoleh, Zoya Bylinskii, Eric Kee, Eli Shechtman, and Yağız Aksoy. 2023. Realistic Saliency Guided Image Enhancement. *Proc. CVPR*.
- S. Mahdi H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yağız Aksoy. 2021. Boosting Monocular Depth Estimation Models to High-Resolution via Content-Adaptive Multi-Resolution Merging. In *Proc. CVPR*.
- Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. 2019. A Multi-Illumination Dataset of Indoor Object Appearance. In *Proc. ICCV*.
- Baptiste Nicolet, Julien Philip, and George Drettakis. 2020. Repurposing a Relighting Network for Realistic Compositions of Captured Scenes. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*.
- Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. 2021. Total Relighting: Learning to Relight Portraits for Background Replacement. *ACM Trans. Graph.* (2021).
- F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. 2016. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *Proc. CVPR*.
- Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A. Efros, and George Drettakis. 2019. Multi-View Relighting Using a Geometry-Aware Network. *ACM Trans. Graph.* 38, 4 (2019).
- Julien Philip, Sébastien Mordant, Michaël Gharbi, and George Drettakis. 2021. Free-viewpoint Indoor Neural Relighting from Multi-view Stereo. *ACM Trans. Graph.* (2021).
- René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* (2020).
- Ke Wang, Michaël Gharbi, He Zhang, Zhihao Xia, and Eli Shechtman. 2023. Semi-supervised Parametric Real-world Image Harmonization. In *Proc. CVPR*.
- Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *Proc. CVPR*.
- Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Binqiang Zhao. 2022. DCCF: Deep Comprehensible Color Filter Learning Framework for High-Resolution Image Harmonization. In *Proc. ECCV*.
- Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. 2022. Learning to Relight Portrait Images via a Virtual Light Stage and Synthetic-to-Real Adaptation. *ACM Trans. Graph.* (2022).
- Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenmann, and Jean-François Lalonde. 2019. All-Weather Deep Outdoor Lighting Estimation. In *Proc. CVPR*.

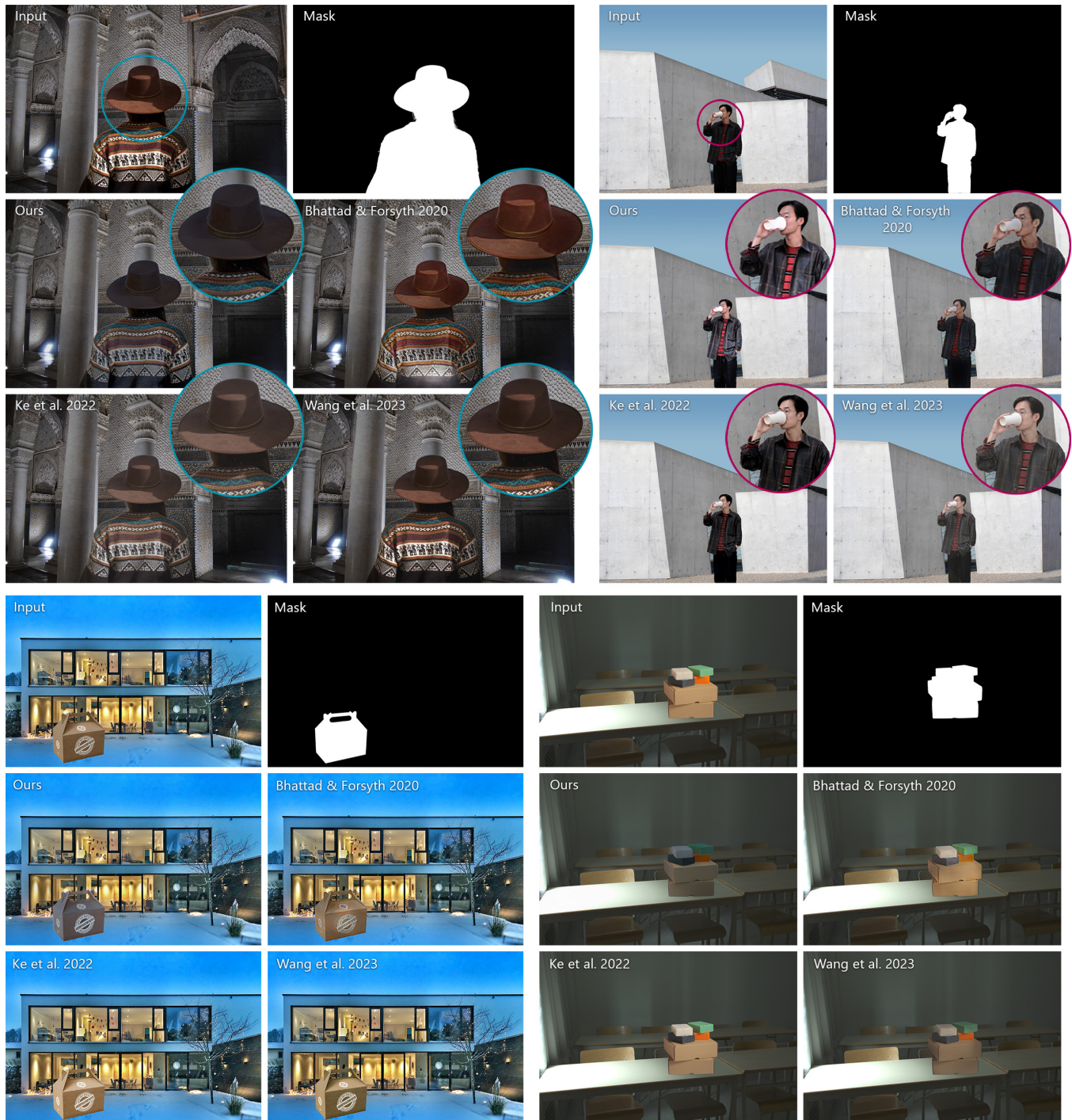


Figure 7: We show comparisons to prior works on difficult composites. Prior methods are not able to fully model the relighting of the foreground, and therefore generate composites with the lighting of the original foreground image. This mismatch results in unrealistic composite images as shown by our user study. Image credit: Unsplash users Stephan Bechert, Grupo Seripafer, JR Harris, victor_g, Jeison Higueta, Miguel Constantin Montes, Jason An and The Laval Indoor Spatially Varying HDR Dataset [Garon et al. 2019]

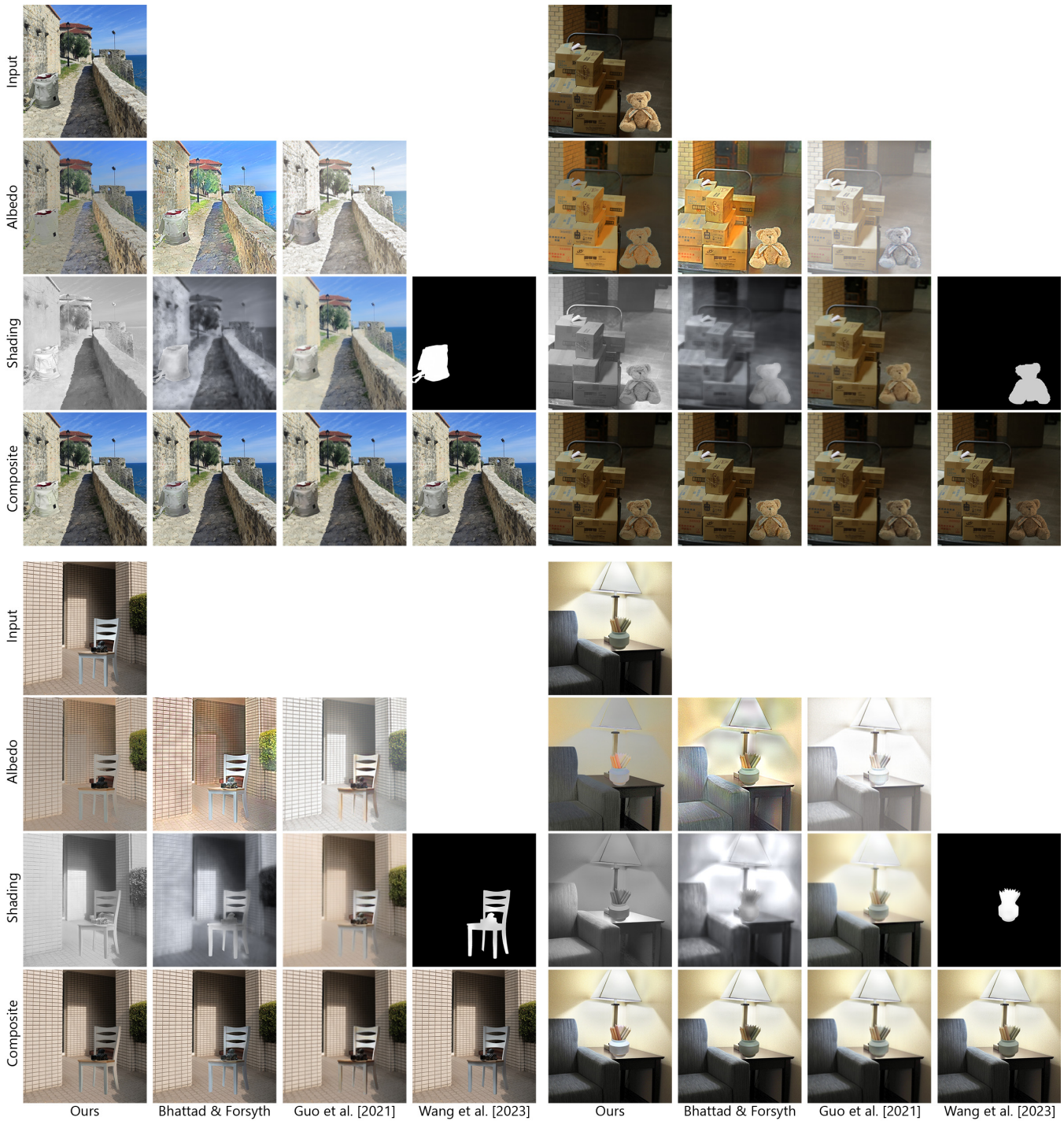


Figure 8: Our method is able to generate physically accurate novel re-shadings under a variety of conditions. Although the works of Guo et al. [2021] and Bhattad and Forsyth [2022] attempt to model the illumination in the form of intrinsic components, they fail to generalize and are not able to estimate meaningful representations. The method of Wang et al. [2023] is able to modulate parametric edits to simulate lighting alteration, their approach doesn't explicitly albedo and shading and therefore cannot generate novel illuminations. Image credit: Unsplash users Laura Ohlman, mak_jp, Solstice Hannan, Randy Fath and Laark Boshoff