Cut-and-Paste Object Insertion by Enabling Deep Image Prior for Reshading

Anand Bhattad D.A. Forsyth
University of Illinois Urbana Champaign
{bhattad2, daf}@illinois.edu

(a) (b) (c) (d)

Figure 1. In two of these images, the spherical lampshade is produced by our method creates realistically shaded renderings from cut-and-paste information; the other two are real photographs. Can you tell which is which? Answer in Sec 4.

Abstract

We show how to insert an object from one image to another and get realistic results in the hard case, where the shading of the inserted object clashes with the shading of the scene. Rendering objects using an illumination model of the scene doesn't work, because doing so requires a geometric and material model of the object, which is hard to recover from a single image. In this paper, we introduce a method that corrects shading inconsistencies of the inserted object without requiring a geometric and physical model or an environment map. Our method uses a deep image prior (DIP), trained to produce reshaded renderings of inserted objects via consistent image decomposition inferential losses. The resulting image from DIP aims to have (a) an albedo similar to the cut-and-paste albedo, (b) a similar shading field to that of the target scene, and (c) a shading that is consistent with the cut-and-paste surface normals. The result is a simple procedure that produces convincing shading of the inserted object. We show the efficacy of our method both qualitatively and quantitatively for several objects with complex surface properties and also on a dataset of spherical lampshades for quantitative evaluation. Our method significantly outperforms an Image Harmonization (IH) baseline for all these objects. They also outperform the cut-and-paste and IH baselines in a user study with over 100 users.

1. Introduction

Inserting objects into images is an appealingly easy rendering paradigm – one just moves objects from one image into another. Applications of this task are abundant, ranging from room planners to image editing for artists to training detectors [30, 9]. But most insertions are not realistic because the shading between the inserted object clashes with the target scene and the object sticks out [21]. Current state-of-the-art (SOTA) methods recover an environment map and render objects using their geometric and physical model [27, 13]. Current single-image methods for shape and material recovery cannot accurately reconstruct realistic objects (say, a lego toy) [29]. Therefore, in this work, we focus on an image-based insertion approach: one takes an object from one image, inserts it into another, and expects a system to correct it. Our method not only synthesizes plausibly realistic renderings of inserted objects with complex surface properties but also does not require a geometric or physical model or an environment map at test time. Furthermore, it does not require rendered images during training.

Our method, DIPR, uses Deep Image Prior [50] for reshading. DIPR adjusts shading so that simple inferences are consistent with cut-and-paste predictions. The rendering process produces an image that is realistic, guaranteed by our use of a deep image prior. The rendering must also produce (a) an albedo that matches the cut-and-paste albedo, (b) a shading that matches the target scene's shading outside the inserted fragment, and (c) the rendered shading and the cut-and-paste normals are consistent with each other. The result is a simple procedure that produces convincing shading. Our entire process, including image decomposition, does not require any form of labeled data for training. We use an extension of Retinex[24], build a statistical process for data generation and train variants of image decomposition models for DIPR. In fact, our only use of simulated ground truth is our use of a pre-trained, off-the-shelf, normal estimator [37].

2022 International Conference on 3D Vision (3DV) | 978-1-6654-5670-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/3DV57658.2022.00045

978-1-6654-5670-8/22/$31.00 ©2022 IEEE
DOI 10.1109/3DV57658.2022.00045

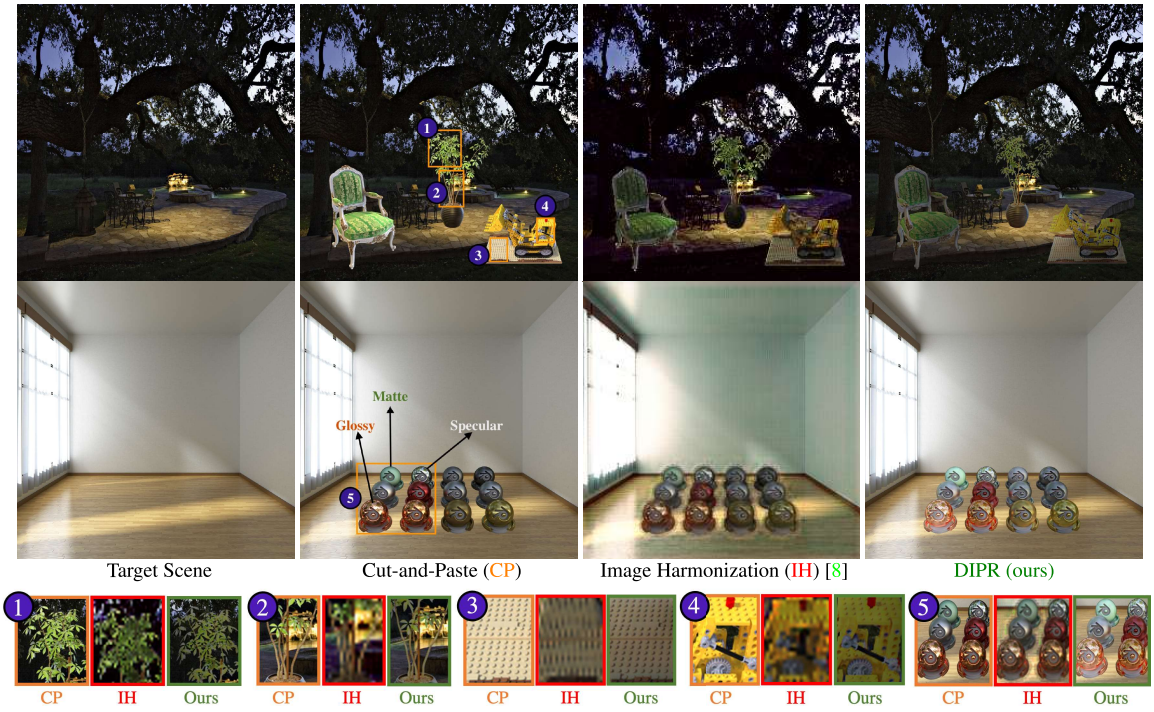


Figure 2. We synthesize realistic, high-resolution renderings of objects added to scenes – cut from one image and pasted into another. Our approach, DIPR, is entirely image-based and can convincingly insert objects with complex surface properties (a lego dozer, a plant and a chair in the first row) and matte, glossy and specular objects (a set of 16 different materials in the second row; crop#5 in bottom row) added to spatially varying illuminating scenes (indoor-outdoor, day-night) without requiring the geometry of the inserted objects or the parameters of the target scene. Key findings from our method are (a) it appears to handle complex and subtle interaction with light; for eg., leaves (crop#1) (b) it appears to understand 3D scene and illumination reasonably well (crop#2); CP in crop#2 looks realistic in local context, but it isn't capturing the 3D scene – the light source behind is far away from the plant, and (c) it preserves material properties because of a carefully designed model; enabling object insertion without any loss of high-frequency details (crop#3 & #4).

Our experiments show DIPR convincingly inserts several objects with complex surface properties – a lego dozer, a plant, a chair and a set of 16 materials with different reflective properties (Fig. 2), cars (Fig. 8) and spherical lampshades (Fig. 7). In qualitative analysis, we show DIPR produces convincing results compared to a SOTA image harmonization baseline [8]. We also find that DIPR achieves significantly better PSNR, MSE and LPIPS scores, outperforming a SOTA image harmonization baseline for lampshades renderings (Tab. 1). We conduct user studies comparing our renderings against baselines and real images. We also show our method has an implicit notion of 3D shape as an emergent property of our consistent reshading (Fig. 11).

In summary, **our main contributions are** (1) we enable deep image priors to explicitly reason about the shading of the scene by a new class of image decomposition model. (2) Our method can realistically insert objects without any ground truth labeled data. The only labeled data that our method requires is a surface normal, obtained from a pretrained network. Other than that, our method is completely self-supervised. (3) Our method works for matte, glossy and specular objects with complex surface properties and without using explicit geometric or physical model of the scene. (4) Our method works for diverse (indoor-outdoor, day-night) spatially varying illuminated complex scenes. (5) Our method produces convincing results compared to a SOTA image harmonization baseline and achieves significantly better PSNR, MSE and LPIPS scores. (6) Our method has an implicit notion of 3D shape as an emergent property of our consistent reshading.

2. Related Work

Object insertion originated with Lalonde *et al.* [21]. They insert objects into target images and control illumination problems by checking objects for compatibility with targets; Bansal *et al.* [1] and Lee *et al.* [26] do so by matching contexts. Poisson blending [39, 18] can resolve nasty boundary artifacts, but significant illumination and color mismatches will cause cross-talk between target and fragment, producing ugly results. Karsch *et al.* [19, 20] how convincing insertions of computer graphics (CG) objects into inverse rendering models. Inverse rendering trained with rendered images can produce excellent reshading of CG objects [41].

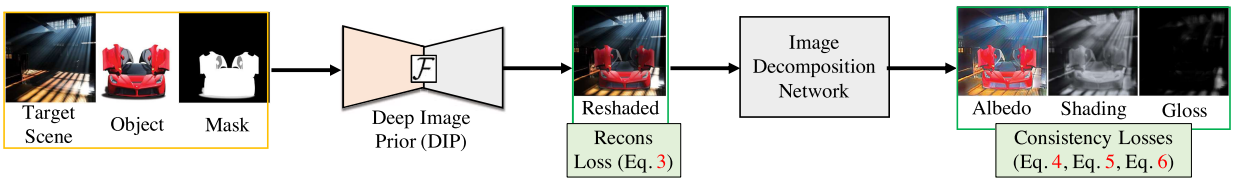


Figure 3. **DIPR overview**. DIPR generates a plausibly realistic rendering of an object inserted from a source image to a target scene. DIPR uses a DIP to generate a reshaded rendering that has consistent image decomposition inferences. The resulting rendering from DIP should have an albedo, same as the cut-and-paste albedo; it should have a shading and gloss field that, outside the inserted fragment, is the same as the target scene's shading and gloss field. The rendering must have similar spherical harmonic properties as target scene and meet a consistency test everywhere (Sec 3.4, Fig. 6). This simple procedure inserts objects convincingly in real images.

However, recovering a renderable model from an image fragment is extremely difficult, particularly if the fragment has an odd surface texture. Liao *et al.*[31, 32] showed that a weak geometric model of the object can be sufficient for correcting shading *if* one has strong the geometric information about the target scene. However, we do not know about geometry of the target scene, except for their normals.

We use **image harmonization** (IH) methods as a strong baseline. IH train models to correct corrupted images where a fragment is adjusted by some noise process (made brighter; recolored; etc.) to the original image [48, 49, 8, 34, 16], and so could clearly be applied here. But we find those IH methods very often change the albedo of an inserted object, rather than its shading. This is because they rely on ensuring consistency of color representations across the image. In contrast, we wish to correct shading alone.

Image Relighting: With appropriate training data, for indoor-scenes, one can predict multiple spherical harmonic components of illumination [13], or parametric lighting model [12] or even full radiance maps at scene points from images [45, 46]. For outdoor scenes, the sun's position is predicted in panoramas using a learning-based approach [17]. However, we do not have access to either training data with lighting parameters/environment maps to construct such a radiance field. Recent single-image relighting methods relight portrait faces under directional lighting [47, 54, 38]. Our method can relight matte, gloss and specular objects with complex material properties like cars (Fig. 8) for both indoor and outdoor spatially varying lighting only from a single image and without requiring physics-based BRDF [27].

Land's Retinex (**image decomposition**) model assumes effective albedo displays sharp, localized changes (which result in large image gradients), and that shading has small gradients [22, 23, 24, 25]. These models require no ground truth. An alternative is to use CG rendered images for training [28, 5, 10]. Current image decomposition evaluation uses the weighted human disagreement rate (WHDR) [3]; current champions are [10, 11]. We use an image decomposition built around approximate statistical models of albedo and shading [11] to train our network without requiring real image decompositions. Our method has reasonable, but not SOTA, WHDR; but we show that improvements in WHDR do not result in improvements in reshading (Fig. 4).

3. Approach

DIPR synthesizes a reshaded object transferred from the source image (s) into a target scene image (t). We use a deep image prior (DIP) [50] as a renderer to produce a reshaded image. We enable DIP to reshade by forcing it to produce consistent image decomposition inferences that meet certain shading consistency tests. We use an image decomposition trained on statistical samples of albedo, shading and gloss; Fig. 5a and not real images (Sec 3.2), and surface normals inferred by the method of [37] to meet the shading consistency tests (Sec 3.4). The final reshaded image's albedo must be like the cut-and-paste albedo; the reshaded image's shading must match the shading of the target scene outside the fragment; and the shading of the reshaded image must have reasonable spherical harmonic properties and meet a consistency test everywhere Fig. 3 summarizes our method.

3.1. Enabling DIP for object reshading

Assume we have a noisy image I_t, and wish to reconstruct the original. Write z for a random vector, and f_θ for a CNN with parameters θ and $E(f_\theta(z); I_t)$ for a loss comparing the image $f_\theta(z)$ to I_t. DIP seeks

$$\hat{\theta} = \text{argmin}_\theta E(f_\theta(z); I_t) \quad (1)$$

and then reports $f_{\hat{\theta}}(z)$. In this naive setup we find that the DIP always converges to cut-and-paste image. This is because the inconsistency we observe in cut-and-paste images are subtle view-dependent lighting effects that are difficult to capture using a simple DIP.

An alternative strategy is to decompose image into two components – a persistent map, one that is invariant to lighting effects and a transient map, one that changes with lighting. We then have to train a DIP only to make corrections to the extrinsic map and use intrinsic map as it is. To this end, we modify Eq. 1 by requiring that $E(\cdot; I_t)$ only to adjust the extrinsic properties of the inserted object. In particular, write g_ϕ for some inference network(s), $t_\psi(I_s, I_t)$ for inferences constructed out of I_t and the source image I_s. DIPR seeks

$$\hat{\theta} = \text{argmin}_\theta E(g_\phi(f_\theta(z)); t_\psi(I_s, I_t)). \quad (2)$$

For us, g_ϕ is an image decomposition network, which is pretrained and fixed.

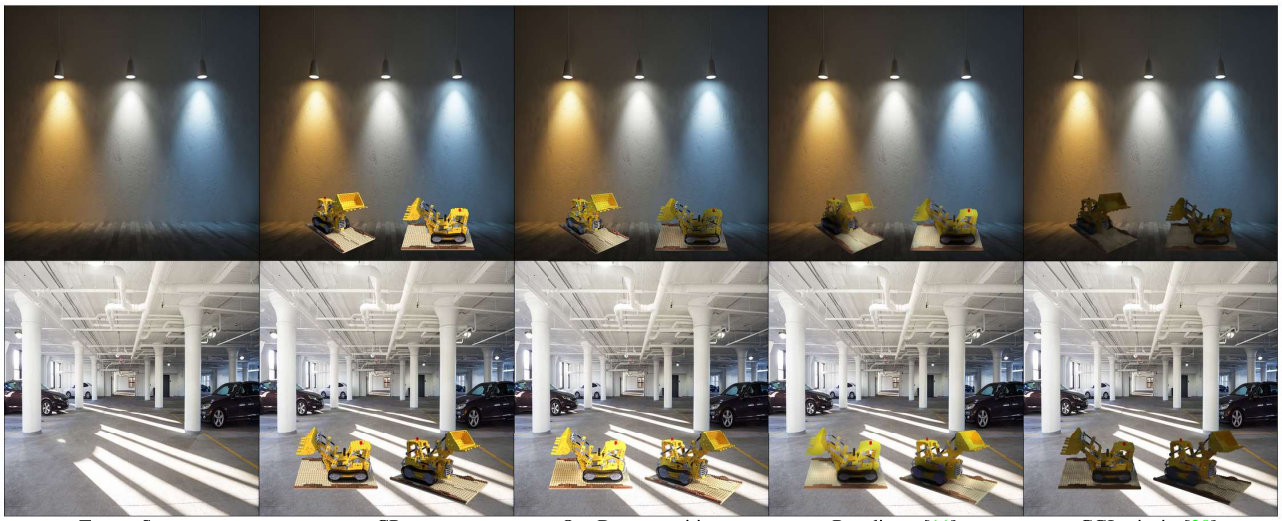


Figure 4. Better WHDR does not mean better reshading. We show reshaded images when using target inferences from different decomposition models. Our decomposition achieves (relatively weak) WHDR of 19%; Paradigms [11] achieve 17%, and a supervised SOTA [28] achieve 15%. Paradigm [11] decomposition produces worse reshading. Moreover, reshading using a supervised SOTA, CGIntrinsics[28], is worse than ours and Paradigms decomposition. This reflects that better recovery of albedo, as measured by WHDR, does not produce better reshading. The key issue is that methods that get low WHDR do so by suppressing small spatial details in the albedo field (for example, the surface detail on the lego dozer), and the shading inference method cannot recover these details, and so they do not appear in the final rendering. From the perspective of reshading, it is better to model them as fine detail in albedo than in shading.

3.2. Image Decomposition

A natural choice for an image decomposition is an albedo map (persistent to lighting) and a shading map (transient to lighting). One could then use a SOTA pretrained image decomposition network as g_ϕ only to adjust the shading of the scene and use the cut-and-paste albedo as it is. Next, we train DIP to reshade (DIPR) the inserted object to produce an image with their albedo, same as the cut-and-paste albedo and a shading field, same as the cut-and-paste image only for the background region other than the inserted fragment. DIPR then learns to extrapolate or interpolate shading for the inserted fragment from the background's shading field.

We first evaluated DIPR with two SOTA image decomposition methods (one supervised [28] and one unsupervised [11]) as measured by strong WHDR performance. The supervised decomposition train their models on CG-generated datasets with ground-truth supervision. [11] uses *Paradigms*, a statistical model of albedo and shading, an extension of the Retinex [24]. Albedo paradigms are Mondrian images. Shading paradigms are Perlin noise [40]. We show these methods result in poor reshading outcomes (Fig 4). SOTA albedo-shading decompositions get strong WHDR performance by suppressing fine spatial details in albedo. These methods preserve spatial and geometric details in the shading field and not albedo because they construct albedo as a piece-wise color constant (Mondrian) with no fine details on them. These fine geometric details are hard to recover accurately from a DIP when trained to interpolate the foreground shading field from their background. However, [11] offers an interesting feature in their Paradigms construction. The statistical models used are authored. Changing the statistical properties of their models would result in a different class of decompositions. We take advantage of this feature. We change [11] and construct an albedo (A) – a persistent map, a diffuse shading (S) – a multiplicative transient map and a gloss (G) – an aditive transient map using the same statistical process. We compose them as $I = A \times S + G$ to form an image and train a network to decompose them back. Fig. 5a shows samples of our. Fig. 5b illustrates the resulting decompositions are satisfactory on MSCOCO [33] real images. The main difference between ours and [11] is that we assume shading to be smooth and albedo has all the high-frequency information so that they can be recovered when reshading with a DIP. This is a reasonable assumption for our method because we aim to preserve all the fine-spatial details of inserted fragment when transferring from one image to another. The additional gloss map that we use helps us to extract better lighting representations in scenes with strong lighting effects like shafts. We show our decomposition produces convincing object reshading when compared to other SOTA image decomposition methods (Fig. 4).

3.3. Base Losses

We first construct the desired target albedo (A_t), target shading and gloss (S_t and G_t). We then train DIPR to produce an image that has reasonable albedo, shading and gloss

Figure 5. **Image decomposition.** Left: samples of our albedo, shading and gloss used to train our image decomposition network. Right: examples showing MS COCO image decompositions.

properties. For DIPR, the input z is the cut-and-paste image and f_θ is optimized to inpaint inserted fragment and also to meet satisfactory image decomposition consistency tests. We use U-Net with partial convolution [35, 44]. However, we find the standard partial convolution converges to a trivial solution, producing images close to cut-and-paste. To prevent this overfitting, we flip the context for partial convolution. We consider the inserted object(s) as the context and hallucinate/outpaint the entire target scene around it. We call this *flipped partial convolution*. This encourages the network not to overfit to the input cut-and-paste image.

We use $CP(I_s; I_t; s)$ for an operator that cuts the fragment out of the source image (I_s), scales it by s, and places it in the relevant location in the target image (I_t). M for a mask with the size of the target image that is 0 inside the fragment and 1 outside. Our reconstruction loss for the background is:

$$\mathcal{L}_{recons} = \|I_t \odot M - (f_\theta(CP(I_s; I_t; s); M))\|^2 \quad (3)$$

We then pass the DIP rendered image through the image decomposition network g_ϕ making A_r, S_r and G_r for the rendered albedo, shading and gloss maps respectively. Our consistent image decomposition inference losses to train DIPR are:

$$\mathcal{L}_{decomp} = \|A_{CP(I_s;I_t;s)} - A_r\|^2 + \|S_t \odot M - S_r \odot M\|^2 + \|G_t \odot M - G_r \odot M\|^2 \quad (4)$$

3.4. Normal Consistency Losses

We use two normal consistency losses to make the strong structure of a shading field apparent to a DIP's reshading. There is good evidence that shading (image extrinsic)is tied across surface normals (this underlies spherical harmonic models [29, 51]), and one should think of a surface normal as a latent variable that explains shading or extrinsic similarities. We assume the resulting illumination approximated with the first 9 spherical harmonics basis coefficients (Y) and does not change when an object is inserted into a scene. We get Y by solving the least square regression between normals (N)

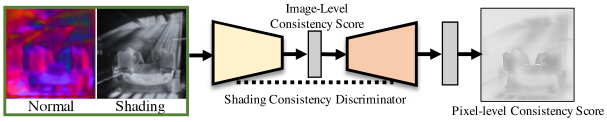


Figure 6. **Shading consistency discriminator** penalizes shading, if it is not consistent with the cut-and-paste normals.

and shading (S) for both the target scene and the resulting composite image. We then minimize loss ($\mathcal{L}_Y$) between the target and rendered image $Y(S; N)$ and use a Huber loss.

$$\mathcal{L}_Y = \begin{cases} \frac{1}{2}(Y(S_t; N_t) - Y(S_r; N_{CP}))^2 & \text{for} |Y_t - Y_r| \leq 1, \\ |Y(S_t; N_t) - Y(S_r; N_{CP})| - \frac{1}{2} & \text{otherwise.} \end{cases} \quad (5)$$

Spherical harmonic shading fields have some disadvantages: every point with the same normal must have the same shading value, which results in poor models of (say) indoor shading on walls. To control this effect, we use a novel neural shading consistency loss ($\mathcal{L}_Z$) that allows the shading field to depart from a spherical harmonic shading field, but only in ways consistent with past inferences. Our shading consistency discriminator, $\zeta(S; N)$, is a U-Net [43] (Fig. 6); trained to discriminate real and fake shading-normal pairs. $\zeta(S; N)$ produces two outputs: one a pixel-level map, yielding the first loss term in Eq. 6, which measures per-pixel consistency; the other an image-level value, the second term in Eq. 6, which measures consistency for the entire image. The $\mathcal{L}_Z$ loss is a binary cross-entropy loss. Let $m \times n$ be the resolution of our renderings, then $\mathcal{L}_Z$ is given by

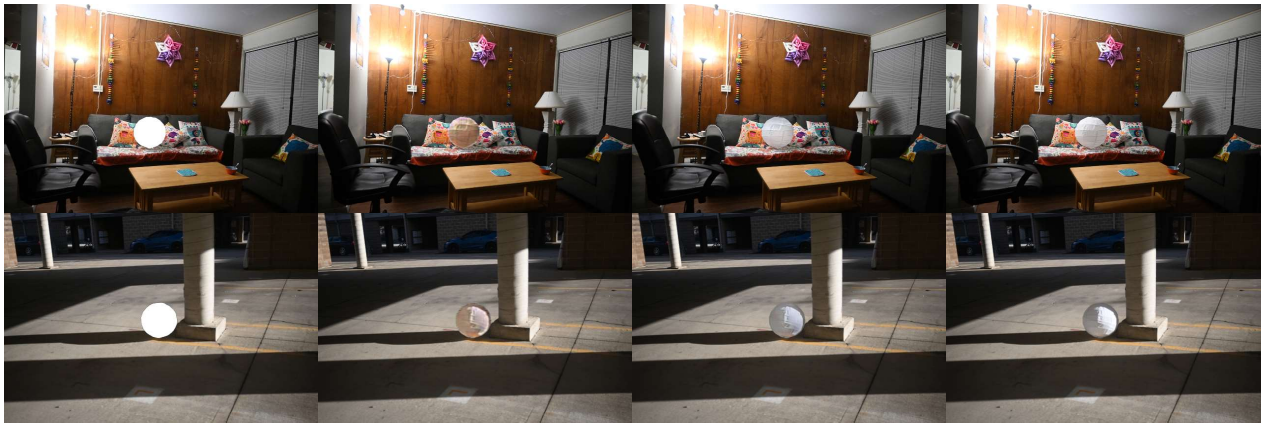
$$\mathcal{L}_Z = -\sum_{i=1}^{m}\sum_{j=1}^{n} \log \zeta(S_r[i,j]; N_{CP}[i,j]) - log\zeta(S_r N_{CP}) \quad (6)$$

In summary, we update DIPR with

$$\mathcal{L}_r = \mathcal{L}_{recons} + \mathcal{L}_{decomp} + \mathcal{L}_Y + \mathcal{L}_Z \quad (7)$$

4. Experiments

Scenes and objects. We collected about 100 diverse images with spatially varying illumination, both indoors-outdoors, day-night, to act as target scenes in our experiments. 25 of these scenes are captured by placing a spherical lampshade, which we use as ground truth for quantitative evaluation. We first test DIPR by inserting simple 2D bright disks at various locations in target scenes (Fig. 10). We also show reshading results for real cars inserted into our target scenes (Fig. 8). We also tested DIPR with complex surface properties – a lego dozer, a plant, a chair, and a set of sixteen materials with different reflective properties used in NeRF [36] (Fig. 2). Other objects from [29] are in our Appendix. We use ADE20K validation set [53] for supplying real residual loss to our image decomposition network and also to train our shading consistency network. ADE20K does not have ground truth normals and we use normals from [37].



CP IH [8] DIPR (ours) Real Scene

Figure 7. **Reshading spherical lampshades.** Similar to Fig. 10, we generate sphere rendering by overlaying 2D white disks over the real sphere from the photograph. We then use DIPR to reshade the overlaid 2D disk and combine that with the actual albedo (as they remain same irrespective of lighting) into a final rendering. IH copies average color. Our renderings are close to the ground truth (also see Tab. 1).

Network architecture. Our DIP network is very similar to the U-Net used in 3DPhotoInpainting [44]. We use their provided model in our implementation as DIP. The only difference is our use of flipped partial convolution instead of standard partial convolution as described in our Sec 3.

Training details. We used U-Net for DIP, Image Decomposition and Shading Consistency Network. Network architecture and other training details are in our supplementary. We update our DIP for a fixed 10k iterations and this takes about 900 seconds using our image decomposition network and 1600 seconds when using CGIntrinsic [28].

IH baseline. We use Cong *et al.*'s DoveNet [8] and their provided pretrained model for our IH baseline.

Quantitative comparison to ground truth. We compare 25 photographs of a spherical lampshade with DIPR based insertions. We get insertions by placing a white disk over the lampshade, reshading the disk using DIPR, then multiplying by lampshade's albedo (a favorable case as it correctly rendered cast shadows). We then quantitatively compare results with ground truth (the real lampshade) using PSNR, LPIPS or MSE (Tab. 1; Fig. 1, Fig. 7). Fig. 1 (b) and (c) are real.

Fooling users. The gold standard evaluation here is user studies (after all, the goal is to fool people). However, user studies are a poor way to polish a method, and a proxy would be valuable. For images of the spherical lampshade used in the quantitative evaluation of Tab. 1, we asked users to identify which of the two presented images (showing distinct scenes, but both containing the lampshade) was real. One image presented was always a rendering, the other always real. Users each see 16 pairs, and there were 72 participants performing this study. Our DIPR renderings are very good at fooling users (50% is a chance). They pick DIPR 42.7% of the time. We then use logistic regression to predict each rendered image's probability of being marked real against MSE, PSNR and LPIPS. An accurate regression would mean that we had a score of "realness". However, such a model explains almost none of the variation of the data (null deviance: 17.87, residual deviance: 17). This means that, while it is pleasant that DIPR has strong MSE, PSNR, and LPIPS scores, these scores can not be used to predict user preferences for our task.

Table 1. **Quantitative Evaluation on Spherical Lampshades** (Fig. 7). Cases: fixed shading fields (S is constant); fixed albedo fields (A constant); cut-and-paste albedo fields ($A_{CP(I_s;I_t;s)}$); image harmonization (IH); and our reconstruction (S_r). Note DIPR reshading wins in all metrics. Note such comparisons to ground truth occur in circumstances favorable to a method like ours (because the shading around the object is consistent), but we know of no way to avoid this.

Shading (S)	Albedo (A)	LPIPS[52] ↓	PSNR ↑
1	1	0.0105	30.81
S_r	1	0.0070	34.26
DoveNet [8]	$A_{CP(I_s;I_t;s)}$	0.0072	35.57
RainNet [34]	$A_{CP(I_s;I_t;s)}$	0.0070	36.10
Harmony Transformer [16]	$A_{CP(I_s;I_t;s)}$	0.0069	36.52
0.25	$A_{CP(I_s;I_t;s)}$	0.0113	26.74
0.50	$A_{CP(I_s;I_t;s)}$	0.0060	31.28
0.75	$A_{CP(I_s;I_t;s)}$	0.0032	38.18
1.0	$A_{CP(I_s;I_t;s)}$	0.0043	36.66
S_r (ours)	$A_{CP(I_s;I_t;s)}$	**0.0021**	**39.53**

Table 2. **Ablation** over losses show each helps improve reshading.

$\mathcal{L}_{decomp}$(Eq. 4)	$\mathcal{L}_{\mathcal{Y}}$ (Eq. 5)	$\mathcal{L}_{\mathcal{Z}}$(Eq. 6)	LPIPS[52] ↓	PSNR ↑
✓			0.0026	37.89
✓		✓	0.0023	38.74
✓	✓		0.0022	39.09
✓	✓	✓	**0.0021**	**39.53**

For other object when we did not have ground truth for objects like cars, lego dozer, plant and chair, we conducted another user study to compare DIPR, CP, and IH renderings. Each study comprises a pre-qualifying process, followed

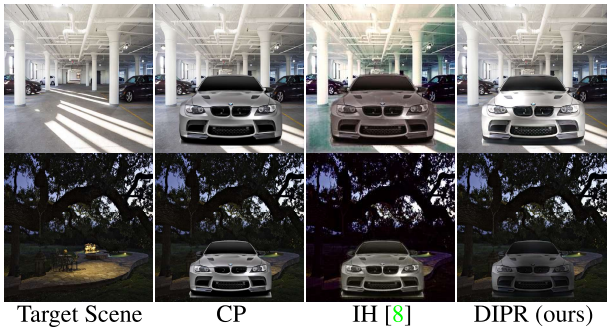


Figure 8. **Reshading real cars.** Glossy effects in car paint with glitter in them make reshading cars a particularly challenging case with their complex reflective properties. DIPR successfully reshades cars without a distinct shift in object and background color produced by IH. Note the bright patch on the metallic bonnet of the grey car in the top row that may possibly be because of the light source just above it.

by 9 pair-wise comparisons, where the user is asked which of two images are more realistic. The result is 109 prequalified studies. The comparisons are balanced. Each study is 3 DIPR-IH pairs, 3 CP-IH pairs, and 3 DIPR-CP pairs, in random order.

We collected data from a total of 122 unique users in 500 studies from Amazon Mechanical Turk. Each study consists of a prequalifying process, followed by 9 pair-wise comparisons, where the user is asked which of two images are more realistic. The prequalifying process presents the user with five tests; each consists of an image with inserted white spheres which are not reshaded (i.e. bright white disks) and an image with inserted spheres which have been reshaded (see Fig 10). We ignore any study where the user does not correctly identify all five reshaded images, on the grounds that the difference is very obvious and the user must not have been paying attention.

The simplest analysis strongly supports DIPR is preferred over both alternatives. One compares the probability that DIPR is preferred to IH (.673, over 327 comparisons, so standard error is .026, and the difference from 0.5 is clearly significant); DIPR is preferred to CP (.645, over 327 comparisons, so the standard error is .026, and the difference from 0.5 is clearly significant); IH is preferred to CP (.511, over 327 comparisons, so standard error is .027, and there is no significant difference from 0.5). An alternative is a Bradley-Terry model [49, 8] used in IH evaluation, regressing the quality predicted by the Bradley-Terry model against the class of algorithm. This yields coefficients of 0 for IH, -0.347 for CP, and 0.039 for DIPR, implying again that DIPR is preferred over IH and strongly preferred over CP.

Consistent Instance Segmentation. We cannot quantitatively evaluate our reshading method when we do not know the ground truth. But we can test whether standard image tasks (which likely benefit from structural consistency in

Figure 9. **Instance segmentation.** DIPR rendered images produce consistent and accurate segmentation maps (bottom row). We observe segmentation fails for cut-and-paste images often. Our key intuition is that the instance segmentation network inherently "knows" if the object's placement is natural or not and expects the foreground object to have consistent shading with the background. If the object's shading does not match with that of the background then the resulting segmentation fails to segment object completely.

Figure 10. **Rendering spheres.** DIPR has some implicit notion of the 3D layout of the scene, which is required to choose the appropriate shading. DIPR shades the white discs as spheres (rather better than IH, implying it "knows" about shape; also see Fig. 11).

images) perform better on our images. We observe image segmentation methods (we used [7]) seem to prefer our images (Fig 9) compared to the naïve cut-and-paste. We believe the instance segmentation network inherently "knows" if the object's placement is natural or not and hence requires the foreground object to be consistent with the background. If not, the segmentation would produce inconsistent results. Our findings are also consistent with Ghiasi *et al.*[15], who show cut-and-paste is a strong data augmentation method for the instance segmentation. Previous models trained without this augmentation, such as [7] are sensitive to cut-and-paste images if the inserted fragments contradict the background's shading. This suggests downstream standard image analysis tasks can serve as a proxy evaluation to further polish reshading methods and also DIPR data augmentations could further improve various recognition tasks.

5. Shape from Shading and its Consistency

Reshadings are derived from consistent shapes. DIPR renderings of circles look like spheres (see Fig 10), suggesting the method has some notion of shape. We test if our shadings are consistent, using two procedures: a large scale using singular values and explicit reconstruction (expensive in compute) at

a small scale. Results suggest there is indeed some emergent notion of shape, obtained with no shape annotated data.

(a) **Large scale:** Imagine we have many different reshadings $S_{i,k}$ of a particular inserted shape (we use white circles in albedo). It is known that multiple shadings of the same geometry have important similarities [2]. Assume that the shading value $S_{i,k}(\mathbf{x})$ is a slowly changing spatial function of the (unknown) normal $\mathbf{N}(\mathbf{x})$, so that $S_{i,k}(\mathbf{x}) = \sum_{m=1}^{N_s} a_{m,k}(\mathbf{x})\phi_m(\mathbf{N}(\mathbf{x}))$. Assume here that k is small, and $a_{m,k}(\mathbf{x})$ are spatially slow. These assumptions apply to, for example, spherical harmonic shading of diffuse surfaces [42]) and shading using the spherical gaussian basis of [27]. Now straighten each image into a vector $\mathbf{S}_{i,k}$. Then these vectors span a N_s dimensional space. Form $\mathcal{D}_i = \left[\mathbf{S}_{i,1}^T, \ldots, \mathbf{S}_{i,N_o}\right]$ (where $N_o > N_s$). We expect the singular values σ_r of $\mathcal{M}_i$ to be small for $r > N_s$ if all the $\mathbf{S}_{i,j}$ are shadings of the same scene, and large if they are not. Using these observations to make a test of consistency requires knowing what is a "small" singular value, and what N_s and N_o should be. We use $N_s = 14$ and 200 sample points on each shading field. We now draw 10000 sets of N_s reshadings from our examples, and look at the test statistic $T = \frac{\Pi_{r=10}^{14}\sigma_r}{\sigma_0^5}$. We see a mean of 0.21 and a standard deviation of 7.4×10^{-2} for T, which appears to be normally distributed. We compare this with a baseline of randomly chosen shading from natural scenes (cropped to the 2D disk). This yields a mean of 0.66 and a standard deviation of 5.9×10^{-2}. We conclude DIPR reshadings of spheres are strongly different from random shading and display a degree of shape consistency.

(b) Explicit shape reconstruction: We produce explicit shape of our inserted circles from their predicted shadings. We take 7 reshaded circles, each from two different images. From this pool of 14, we draw 7 at random, use them to drive a shape reconstruction procedure (see Appendix; a typical reconstruction in Fig. 11). We then compute the spherical harmonic reshading of the resulting reconstruction that is closest to each of the 7 held out images and record the mean squared error of the shading residual for each. A small residual means that the shape is consistent across the reshaded circles. Box plots of the resulting residuals for 7 different splits of the data in Fig. 11. To calibrate, we compare this with 3 baselines; we reshade: (1) a sphere using actual spherical harmonic shading, reconstruct from reshadings, then predict the held out reshadings ('SphHarm'), (2) a constant height surface ('Const') and (3) a surface reconstructed from smoothed random noise shadings ('Rand'). These residuals and singular value analysis suggest the reshading network has some form of shape theory as an emergent property.

6. Discussion and Conclusions

Limitations. DIPR is slow because DIP takes 15 mins to render. DIPR cannot cast shadows; very hard and we leave

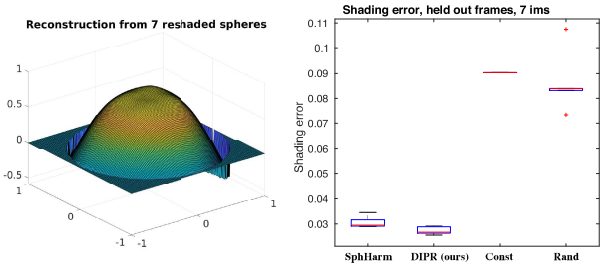


Figure 11. **Shape theory and shading consistency:** On the **left** a sample surface reconstruction produced by using 7 reshaded spheres (Fig. 10). On the right, predicted shading residuals for held out spheres' shadings using our reconstructed shapes. The residuals suggest: there is a consistent underlying shape. Our DIPR residuals are small. The reconstruction process is reliable but the underlying surface is not quite a sphere. 'SphHarm' residuals are also small, but not as small as DIPR. 'Const' residuals are large. Therefore, underlying surface is not flat. Smooth 'Rand' residuals are large, that is, random shadings cannot explain this consistency.

CP DIPR (ours) CP DIPR (ours)

Figure 12. **Failure cases.** Red arrows point to shading errors. In first image, DIPR aggressively copies background shading onto the chair. However, lego's and plant's shading looks plausible. In the second scene, it has two dominant normals – the ground (upwards) and the sky (towards viewer). The lack of third direction results in copying shading either from the sky or the ground.

that for our future work. DIPR likely copies shading, so has problems when there is little shading variation and responds poorly when there are "few" normals in the scene (Fig. 12). **Why DIPR works?** Corrections to object shading cannot be veridical. [31, 32] finds corrected shading often fool humans more effectively than physically accurate lighting, likely because humans attend to complex materials much more than to consistent lighting [4]. The alternative physics theory [6] argues that the brain employs a set of rules that are convenient, but not strictly physical and a violation leads to perception alarm or affects recognition negatively [14]. Otherwise, the scene "looks right". This means humans may tolerate a fair degree of error, as long as it is of the right kind. By requiring image to produce consistent inferences, we appear to be forcing errors to be "of the right kind".

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. 2106825 and 1718221. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Aayush Bansal, Yaser Sheikh, and Deva Ramanan. Shapes and context: In-the-wild image synthesis & manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2317–2326, 2019. 2

[2] Peter N Belhumeur and David J Kriegman. What is the set of images of an object under all possible illumination conditions? *International Journal of Computer Vision*, 1998. 8

[3] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 2014. 3

[4] Julia Berzhanskaya, Gurumurthy Swaminathan, Jacob Beck, and Ennio Mingolla. Remote effects of highlights on gloss perception. *Perception*, 2005. 8

[5] Sai Bi, Xiaoguang Han, and Yizhou Yu. An l 1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM Transactions on Graphics (TOG)*, 2015. 3

[6] Patrick Cavanagh and George A Alvarez. Tracking multiple targets with multifocal attention. *Trends in cognitive sciences*, 9(7):349–354, 2005. 8

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 7

[8] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8394–8403, 2020. 2, 3, 6, 7

[9] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1301–1310, 2017. 1

[10] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 3

[11] DA Forsyth and Jason J Rock. Intrinsic image decomposition using paradigms. *arXiv preprint arXiv:2011.10512*, 2020. 3, 4

[12] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3

[13] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast spatially-varying indoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 3

[14] Isabel Gauthier, Pepper Williams, Michael J Tarr, and James Tanaka. Training 'greeble'experts: a framework for studying expert object recognition processes. *Vision research*, 38(15-16):2401–2428, 1998. 8

[15] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. *arXiv preprint arXiv:2012.07177*, 2020. 7

[16] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14870–14879, 2021. 3, 6

[17] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

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

[19] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 2011. 2

[20] Kevin Karsch, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fonte, Michael Sittig, and David Forsyth. Automatic scene inference for 3d object compositing. *ACM Transactions on Graphics (TOG)*, 2014. 2

[21] Jean-François Lalonde, Derek Hoiem, Alexei A Efros, Carsten Rother, John Winn, and Antonio Criminisi. Photo clip art. *ACM transactions on graphics (TOG)*, 2007. 1, 2

[22] Edwin H Land. Color vision and the natural image. part i. *Proceedings of the National Academy of Sciences of the United States of America*, 1959. 3

[23] Edwin H Land. Color vision and the natural image part ii. *Proceedings of the National Academy of Sciences of the United States of America*, 1959. 3

[24] Edwin H Land. The retinex theory of color vision. *Scientific american*, 1977. 1, 3, 4

[25] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 1971. 3

[26] Donghoon Lee, Sifei Liu, Jinwei Gu, Ming Yu Liu, Ming Hsuan Yang, and Jan Kautz. Context-aware synthesis and placement of object instances. In *Advances in Neural Information Processing Systems*, 2018. 2

[27] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 3, 8

[28] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3, 4, 6

[29] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG)*, 2018. 1, 5

[30] Zicheng Liao, Ali Farhadi, Yang Wang, Ian Endres, and David Forsyth. Building a dictionary of image fragments. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3442–3449. IEEE, 2012. 1

[31] Zicheng Liao, Kevin Karsch, and David Forsyth. An approximate shading model for object relighting. In *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3, 8

[32] Zicheng Liao, Kevin Karsch, Hongyi Zhang, and David Forsyth. An approximate shading model with detail decomposition for object relighting. *International Journal of Computer Vision*, 2019. 3, 8

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 2014. 4

[34] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3, 6

[35] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5

[36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *arXiv preprint arXiv:2003.08934*, 2020. 5

[37] Vladimir Nekrasov, Thanuja Dharmasiri, Andrew Spek, Tom Drummond, Chunhua Shen, and Ian Reid. Real-time joint semantic segmentation and depth estimation using asymmetric annotations. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019. 1, 3, 5

[38] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, Epic Games, Andreas Lehrmann, and AI Borealis. Learning physics-guided face relighting under directional light. 2020. 3

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

[40] Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 1985. 4

[41] Vilayanur S Ramachandran. Perceiving shape from shading. *Scientific American*, 1988. 2

[42] Ravi Ramamoorthi and Pat Hanrahan. On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object. *JOSA A*, 2001. 8

[43] Edgar Schönfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. *arXiv preprint arXiv:2002.12655*, 2020. 5

[44] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020. 5, 6

[45] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[46] Pratul P Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[47] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (Proceedings SIGGRAPH)*, 2019. 3

[48] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM Transactions on Graphics (TOG)*, 29(4):1–10, 2010. 3

[49] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3789–3797, 2017. 3, 7

[50] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 3

[51] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 5

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[53] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 5

[54] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3