CUT-AND-PASTE NEURAL RENDERING

Anonymous authors
Paper under double-blind review

Target Scene | Cut-and-Paste | Cong et al. (2020) | Ours

Figure 1: We generate realistic renderings of cut-and-paste images. Our method is entirely image-based and can convincingly reshade/relight fragments with complex surface properties (a lego dozer, a plant and a chair in top row) and matte, glossy and specular fragments (a set of 16 different materials in bottom row) added to a spatially-varying illuminated target scene (indoor-outdoor and day-night) without requiring the geometry of the inserted fragment or the parameters of the target scene.

ABSTRACT

Cut-and-paste methods take an object from one image and insert it into another. Doing so often results in unrealistic looking images because the inserted object's shading is inconsistent with the target scene's shading. Existing reshading methods require a geometric and physical model of the inserted object, which is then rendered using environment parameters. Accurately constructing such a model only from a single image is beyond the current understanding of computer vision.

We describe an alternative procedure – cut-and-paste neural rendering, to render the inserted fragment's shading field consistent with the target scene. We use a Deep Image Prior (DIP) as a neural renderer trained to render an image with consistent image decomposition inferences. The resulting rendering from DIP should have an albedo consistent with composite albedo; it should have a shading field that, outside the inserted fragment, is the same as the target scene's shading field; and composite surface normals are consistent with the final rendering's shading field. The result is a simple procedure that produces convincing and realistic shading. Moreover, our procedure does not require rendered images or image-decomposition from real images in the training or labeled annotations. In fact, our only use of simulated ground truth is our use of a pre-trained normal estimator. Qualitative results are strong, supported by a user study comparing against state-of-the-art image harmonization baseline.

1 INTRODUCTION

Cut-and-Paste rendering involves creating a new image by cutting fragments out of one or more source images and pasting them into a target image; the idea originates with Lalonde et al. (2007). Results are often unrealistic, because of the difference in illumination between the source and target images. But the procedure is useful to artists, and there is consistent evidence that such procedures can be used to train detectors (Liao et al., 2012; Dwibedi et al., 2017). When the geometry and material of the inserted object are known, it is enough to infer an illumination model from the target, render and composite. But current procedures for recovering shape and material from a single fragment simply can't deal with most realistic fragments (think of, say, a furry cat).

This paper describes a new method, *Cut-and-Paste Neural Rendering*, that can render convincing composite images by adjusting the cut-and-paste images so that some simple image inferences are consistent with cut-and-paste predictions. So the albedo from the adjusted image should look like cut-and-paste albedo; the shading should look like a shading field; and the image should look like an image. A simple post processing trick produces very high resolution composites. Evaluation is mostly qualitative, but we show that our method fools a recent method for detecting tampering.

Our contribution is a method that can realistically correct shading in composite images, without requiring labelled data; our method works for matte, glossy and specular fragments without an explicit geometric or physical model; and human subjects prefer the results of our method over cut-and-paste and image harmonization.

2 RELATED WORK

Object Insertion starts with Lalonde *et. al.* Lalonde et al. (2007), who insert fragments into target images. (Lalonde et al., 2007) control illumination problems by checking fragments for compatibility with targets; Bansal et al. (2019) do so by matching contexts. Poisson blending (Pérez et al., 2003; Jia et al., 2006) can resolve nasty boundary artifacts, but significant illumination and color mismatches will result in cross-talk between target and fragment, producing ugly results. Karsch *et. al.* show that computer graphics objects can be convincingly inserted into inverse rendering models obtained with geometric inference (Karsch et al., 2011) or with single image depth reconstruction (Karsch et al., 2014). Inverse rendering trained with rendered images can produce excellent reshading of computer graphics objects Ramachandran (1988). However, recovering a renderable model from an image fragment is extremely difficult, particularly if the fragment has odd surface texture. Liao *et al.* showed that a weak geometric model of the fragment can be sufficient to correct shading, if one has strong geometric information about the target scene Liao et al. (2015; 2019). In contrast, our work is entirely image based: one takes a fragment from one image, drops it into another, and expects a system to correct it.

We use **image harmonization** methods as a strong baseline. These procedures aim to correct corrupted images. Methods are trained to correct images where a fragment has been adjusted by some noise process (made brighter; recolored; etc.) to the original image (Sunkavalli et al., 2010; Tsai et al., 2017; Cong et al., 2020), and so could clearly be applied here. But we find that image harmonization 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. For example in the iHarmony dataset from Cong *et al*, pink candy is changed to brown (an albedo change; see Fig 12 in Appendix). 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 Garon et al. (2019), or parametric lighting model (Gardner et al., 2019) or even full radiance maps at scene points from images (Song & Funkhouser, 2019; Srinivasan et al., 2020). For outdoor scenes, the sun's position is predicted in panoramas using a learning-based approach (Hold-Geoffroy et al., 2019). One can also construct a volumetric radiance field from multi-view data to synthesize novel views (Mildenhall et al., 2020). However, we do not have access to either training data with lighting parameters/environment maps or multi-view data to construct such a radiance field. Our renderings are entirely image-based. Recent single-image relighting methods relight portrait faces under directional lighting (Sun et al., 2019; Zhou et al., 2019; Nestmeyer et al., 2020). Our method can relight matte, gloss and specular objects with

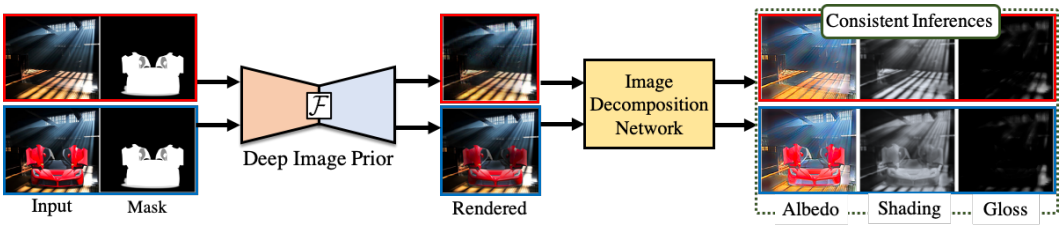


Figure 2: **Cut-and-Paste Neural Renderer**. Given two images, a target scene (top row) and a new fragment added to this target scene (bottom row), our approach generates a plausible, realistic rendering of the composite image by correcting the fragment's shading. Our method uses DIP as a neural renderer trained to produce 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. This simple process produces convincing composite rendering for any cut-and-paste images.

complex material properties like cars (Fig 7) for both indoor and outdoor spatially-varying illuminated environments only from a single image and without requiring physics-based BRDF (Li et al., 2020).

Image decomposition. Land's influential Retinex model assumes effective albedo displays sharp, localized changes (which result in large image gradients), and that shading has small gradients (Land, 1959a;b; 1977; Land & McCann, 1971). No ground truth is required by these models. An alternative is to use CG rendered images for image decomposition training (Li & Snavely, 2018), particularly with specialized losses (Bi et al., 2015; Fan et al., 2018). One can also train using rendering constraints to produce a form of self-supervised training (Janner et al., 2017). Current image decomposition evaluation uses the weighted human disagreement rate (WHDR) (Bell et al., 2014); current champions are (Fan et al., 2018). We use an image decomposition method built around approximate statistical models of albedo and shading (*paradigms*) to train our image decomposition network without requiring real image ground truth decompositions. Our method has reasonable, but not SOTA, WHDR; but we show that improvements in WHDR do not result in improvements in reshading (Fig 5)

3 Cut-and-Paste Neural Renderer

We synthesize a reshaded composite image containing a fragment transferred from a source image into a target scene image. We use a deep image prior (DIP) (Ulyanov et al., 2018) as a neural renderer to produce a reshaded image that produces consistent image decomposition inferences. We use an image decomposition trained on Paradigms (statistical samples of albedo, shading and gloss) and not real images described in section 3.3, and normals inferred by the method of Nekrasov et al. (2019) to meet the shading consistency tests (section 3.2). 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 Figure 2 summarizes our method.

3.1 Deep Image Prior for Rendering Cut-and-Paste Images

Assume we have a noisy image $\mathcal{I}_t$, and wish to reconstruct the original. Write z for a random vector, and f_θ for a CNN with parameters θ and $E(f_\theta(z); \mathcal{I}_t)$ for a loss comparing the image $f_\theta(z)$ to $\mathcal{I}_t$. The Deep Image Prior seeks

$$\hat{\theta} = \text{argmin}_\theta E(f_\theta(z); \mathcal{I}_t) \tag{1}$$

and then reports $f_{\hat{\theta}}(z)$. We modify this formulation by requiring that the $E(\cdot; \mathcal{I}_t)$ impose inferential consistency. In particular, write g_ϕ for some inference network(s) and $t_\psi(\mathcal{I}_s, \mathcal{I}_t)$ for inferences constructed out of $\mathcal{I}_t$ and the source image $\mathcal{I}_s$. We seek

$$\hat{\theta} = \text{argmin}_\theta E(g_\phi(f_\theta(z)); t_\psi(\mathcal{I}_s, \mathcal{I}_t)). \tag{2}$$

For us, g_ϕ is an image decomposition network (pretrained and fixed), and t_ψ creates target albedos ($\mathcal{A}_t$), shading ($\mathcal{S}_t$) and glosses ($\mathcal{G}_t$) fields. DIP is then trained to produce an image that has reasonable

intrinsic image properties. For DIP, the input z is the cut-and-paste image and f_θ is optimized to inpaint inserted fragment and also to meet satisfactory intrinsic image properties.

We use a U-Net with partial convolutions and skip connections (Liu et al., 2018). However, we find the standard partial convolution often convergence to a trivial solution producing images that are close to cut-and-paste. To prevent this overfitting to cut-and-paste images, we switch the context for the partial convolution; that is we consider inserted fragment as the context and hallucinate/outpaint the entire target scene around it. This can be viewed as a *inverse partial convolution*.

We use $\mathcal{CP}(\mathcal{I}_s; \mathcal{I}_t; s)$ for an operator that cuts the fragment out of the source image $\mathcal{I}_s$, scales it by s, and places it in the relevant location in the target image $\mathcal{I}_t$. $\mathcal{M}$ for a mask with size of the target image that is 0 inside the fragment and 1 outside. Reconstruction loss for background and is given by:

$$\mathcal{L}_{recons} = \|\mathcal{I}_t \odot \mathcal{M} - (f_\theta(\mathcal{CP}(\mathcal{I}_s; \mathcal{I}_t; s); \mathcal{M}))\|^2 \tag{3}$$

The DIP rendered image is then passed through the image decomposition network g_ϕ making $\mathcal{A}_{render}$, $\mathcal{S}_{render}$ and $\mathcal{G}_{render}$ for the albedo, shading and gloss respectively. Our consistent image decomposition losses to train DIP are:

$$\mathcal{L}_{decomp} = \|\mathcal{A}_{\mathcal{CP}(\mathcal{I}_s;\mathcal{I}_t;s)} - \mathcal{A}_{render}\|^2 + \|\mathcal{S}_t \odot \mathcal{M} - \mathcal{S}_{render} \odot \mathcal{M}\|^2 + \|\mathcal{G}_t \odot \mathcal{M} - \mathcal{G}_{render} \odot \mathcal{M}\|^2 \tag{4}$$

3.2 Shading Consistency Losses

We use two shading consistency losses to make the very strong structure of a shading field apparent to a DIP's rendering. There is good evidence that shading is tied across surface normals (this underlies spherical harmonic models (Liao et al., 2019; Li et al., 2018; Yu & Smith, 2019)), and one should think of a surface normal as a latent variable that explains shading similarities in images. We assume that the resulting illumination, approximated with the first 9 spherical harmonics coefficients (SHC; a 9 dimensional vector) do not change when new objects are added to a target scene. We obtain SHC by solving the least square regression between normals ($\mathcal{N}$) and shading for both the target and the resulting rendered image. We get normal estimates from Nekrasov et al. (2019) and then minimize loss ($\mathcal{L}_{SHC}$) between the target and rendered image's SHC($\mathcal{S}; \mathcal{N}$). We use Huber loss for $\mathcal{L}_{SHC}$.

$$\mathcal{L}_{SHC} = \begin{cases} 0.5 \times (SHC(\mathcal{S}_t; \mathcal{N}_t) - SHC(\mathcal{S}_{render}; \mathcal{N}_{\mathcal{CP}}))^2 & \text{for} |SHC_t - SHC_{render}| \leq \mathbf{1}, \\ |SHC(\mathcal{S}_t; \mathcal{N}_t) - SHC(\mathcal{S}_{render}; \mathcal{N}_{\mathcal{CP}})| - 0.5 & \text{otherwise.} \end{cases} \tag{5}$$

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 neural shading consistency loss ($\mathcal{L}_{NSC}$) that allows the shading field to depart from a spherical harmonic shading field, but only in ways that are consistent with past inferences. Inspired from a pixel-wise discriminator in Schönfeld et al. (2020), we use a shading consistency U-Net; $\zeta(\mathcal{S}; \mathcal{N})$, trained to discriminate real and fake shading-normal pairs. The $\mathcal{L}_{NSC}$ loss is a binary cross-entropy loss for each pixel pairs and also a global score at Encoder's code level to discriminate if the pair is real (consistent) or fake (inconsistent). We train DIP to produce a shading filed that produces pixel-wise consistent outcome with respect to the normals. Let $m \times n$ be the resolution of image, then $\mathcal{L}_{NSC}$ can be simplified into

$$\mathcal{L}_{NSC} = -\sum_{i=1}^{m}\sum_{j=1}^{n} \log \zeta(\mathcal{S}_{render}[i,j]; \mathcal{N}_{\mathcal{CP}}[i,j]) - log\zeta(\mathcal{S}_{render}\mathcal{N}_{\mathcal{CP}}) \tag{6}$$

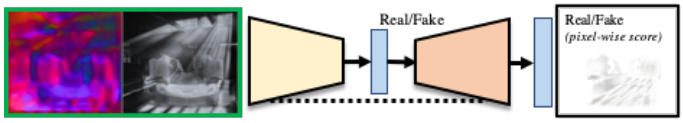


Figure 3: **Shading Consistency U-Net** is trained separately to discriminate consistent and inconsistent pixel-level shading-normal pairs.

In summary, our overall rendering losses that we minimize to update DIP is

$$\mathcal{L}_{render} = \mathcal{L}_{recons} + \mathcal{L}_{decomp} + \mathcal{L}_{SHC} + \mathcal{L}_{NSC} \tag{7}$$

3.3 Intrinsic Image Decomposition

We believe any competent image decomposition network could be used. However, our experience suggests accurate albedo recovery (measured by strong WHDR performance) results in poor reshading

Figure 4: **Image Decomposition.** Left: samples drawn from our *paradigms* that are used to train our image decomposition network. Right: examples showing MS COCO image decompositions after training on paradigms. Note our image decomposition network is trained without real image labeled data.

outcomes (Fig 5).It is also helpful to have a small and efficient network to reduce the overall back-propagation time when training a DIP. Therefore, we trained a small U-Net to produce albedo, shading and gloss layers from images using **paradigms**, samples from statistical models intended to capture the spatial statistics of albedo, gloss and shading. This is a simple extension of the Retinex model (Land, 1977). Albedo paradigms are Mondrian images. Shading paradigms are Perlin noise (Perlin, 1985), and gloss paradigms consist of light bars on a dark background. These are used to produce fake images by a simple composition ($\mathcal{AS} + \mathcal{G}$) that in turn are used to train image decomposition network. Figure 4a shows some samples from each. As Figure 4b illustrates, the resulting intrinsic image models are satisfactory on MSCOCO real images (Lin et al., 2014). For our experiments, we trained two models (*Paradigm I* and *Paradigm II)*; samples drawn from different statistical models, to investigate the consequences of strong WHDR recovery (II has significantly better WHDR than I). The main difference between the two is that Paradigm I has high-frequency, fine-grained details in albedo and not shading, and for Paradigm II it is opposite.

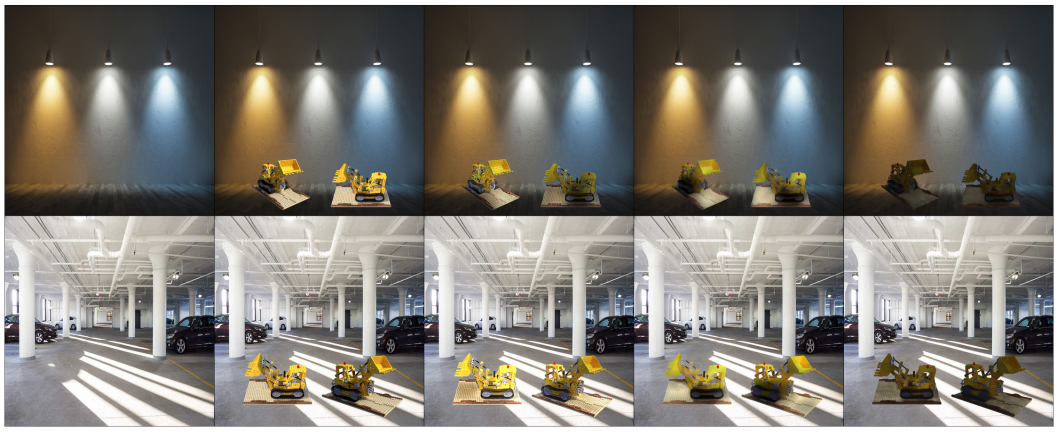


Figure 5: Better WHDR does not mean better reshading. Reshaded images when using target inferences from different image decomposition networks. Paradigm I achieves (relatively weak) WHDR of 22%; Paradigm II achieves 19%, close to state-of-the-art (SOTA) for an image decomposition network that does not see rendered images (Liu et al., 2020). However, Paradigm II produces significantly worse reshading results. Furthermore, reshading using CGIntrinsics (Li & Snavely, 2018) (a supervised SOTA method) is also qualitatively worse than using Paradigm I. 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), but the normal inference method cannot recover these details, and so they do not appear in the reshaded image. From the perspective of reshading, it is better to model them as fine detail in albedo than as surface normals.

Figure 6: Our method has some implicit notion of the 3D layout of the scene, which is required to choose the appropriate shading. Our method shades the white discs as spheres (rather better than IH, implying it "knows" about shape). Furthermore, a depth network Ranftl et al. (2020) applied to our reshaded results recovers depths that span the volume of the scene rather more than for others. Note the curious fact that the depth network can choose depths for CP and IH images confidently, too.

3.4 POST-PROCESSING

Removing DIP artifacts. In our approach (see Fig 2), DIP sees both target image and cut-and-paste image. We require that DIP inpaints the shading field of the original target image *and* also the cut-and-paste composite image. This process is analogous to rendering scenes once with an object added and once without an object (see (Debevec, 1998; Karsch et al., 2011)). Doing so means the target scene's shading field acts like a regularizer that prevents DIP copying cut-and-paste. Furthermore, we can remove DIP specific noisy artifacts from the reconstructed image. Using the notation above and writing $\mathcal{I}_{obj}$ for the target image rendered by DIP with the object and $\mathcal{I}_{noobj}$ for the target scene rendered by DIP (i.e. no object), we form

$$\mathcal{I}_{final} = (1 - \mathcal{M}) \odot \mathcal{I}_{obj} + \mathcal{M} \odot (\mathcal{I}_t + \mathcal{I}_{obj} - \mathcal{I}_{noobj}) \tag{8}$$

High-Res Rendering. Our image decomposition network can decompose very high resolution albedos reliably. Since our shading and gloss are both locally smooth for Paradigm I, we can easily upsample them to high resolution. This allows us to render very-high-resolution final reshaded images (1024p) without any additional computational budget. Note that we train DIP to render images with 256p resolution only.

4 EXPERIMENTS

Dataset. We collected about 75 diverse images with spatially-varying illumination, both indoors and outdoors to act as target scenes in our experiments. We tested our cut-and-paste rendering on circle cutouts as spheres (Fig 6), ficus, chair, lego and a set of sixteen materials used in NeRF (Mildenhall et al., 2020) (Fig. 8). We also show reshading results for cars (Fig 7). Many other objects can be found in our Appendix. We use ADE20k validation set Zhou et al. (2016) for supplying real residual loss to our image decomposition network and also to train our shading consistency network. Note that the ADE20k does not have ground truth surface normals and we use normals from a pretrained network Nekrasov et al. (2019) as described in section 3.2.

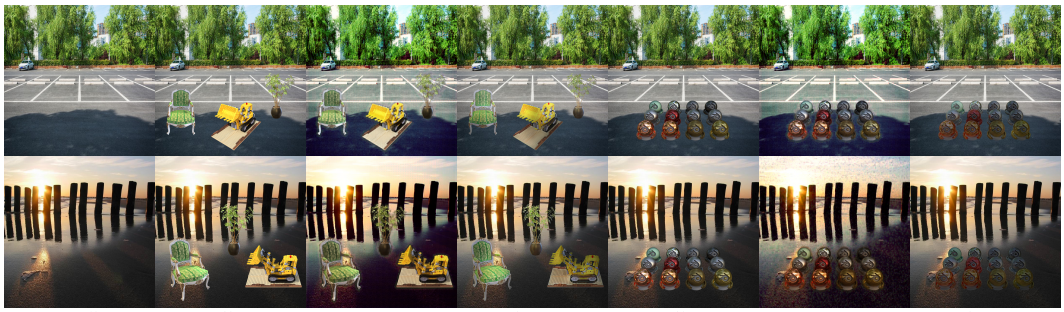
Training Details. We used U-Net for DIP, Image Decomposition and Shading Consistency Network. Network architecture and other training details are in our supplementary. We update our DIP is updated for a fixed 10k iterations and this takes about 1000 seconds using our image decomposition network and 1600 seconds when using CGIntrinsic (Li & Snavely, 2018).

Intrinsic images: Figure 5 explains how better albedo recovery can lead to weaker reshading. Paradigm I achieves realistic reshading and we use image decomposition model trained on Paradigm I for all our experiments.

Fooling users. We wish to know whether our reshaded images are reliably preferred to cut and paste images. We have conducted a user study to compare our reshading method (RS), cut-and-paste (CP), and image harmonization (IH). We collected data from a total of 122 unique users, in 500 studies.

Target Scene CP IH Ours CP IH Ours

Figure 7: Glossy effects in car paint (with glitter in them) make cars a particularly challenging case with their complex reflective properties. Our method successfully reshades cars without the distinct shift in object and background color produced by IH.



Target Scene CP IH Ours CP IH Ours

Figure 8: Our method reshades objects with complex surface geometry and complex material properties, and appears to preserve the identity of the material without the distinct shift in object and background color produced by IH.

Each study consists of a prequalifying process, followed by 9 pairwise 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. 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 result is 109 prequalified studies. The comparisons are balanced (each study is 3 RS-IH pairs, 3 CP-IH pairs and 3 RS-CP pairs, in random order and presentation).

The simplest analysis strongly supports the idea that RS is preferred over both alternatives. One compares the probability that RS is preferred to IH (.673, over 327 comparisons, so standard error is .026, and the difference from 0.5 is clearly significant); RS is preferred to CP (.645, over 327 comparisons, so 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 (Tsai et al., 2017; Cong et al., 2020) used in image harmonization 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 RS, implying again that RS is preferred over IH and strongly preferred over CP.

Quantitative Analysis. We cannot quantitatively evaluate our reshading method, because we do not know ground truth. But we tried scoring images based on a quality when no reference is available (Mittal et al., 2012). Quite surprisingly, NIQE gets significantly worse score for target scenes without any fragments added. This shows that getting a reliable quantitative evaluations for reshading is extremely hard. We also tried the recent CNN detector of image fakes described by Wang et al. (2020). We find that it does not detect any of our synthesized images as fake; but it also fails to detect cut-and-paste images and image harmonization images.

Target Scene | CP | IH | $\mathcal{L}_{recons} + \mathcal{L}_{decomp}$ | $+\mathcal{L}_{SNC}$ | $+\mathcal{L}_{SNC} + \mathcal{L}_{SHC}$

Figure 9: **Ablation study** showing how different components of our loss helps in correcting overall shading. Results improve moving from left to right. We observe that our $\mathcal{L}_{SNC}$ helps in improving local changes based on the near-by surroundings and $\mathcal{L}_{SHC}$ takes into account the overall illumination and hence providing complementary solutions.

Images	NIQE↓	CNN-D↓	B-T↑
IH	38.39	0	0
CP	**9.67**	0	-0.35
Ours	9.78	0	**0.04**
Target Scene	12.51	0	NA

Table 1: Quantitative Analysis without ground truth is hard and unreliable. We report numbers obtained from a no reference image quality evaluator –NIQE (Mittal et al., 2012), a CNN based tampering detector – CNN-D (Wang et al., 2020) and Bradley-Terry model which is used in image harmonization (IH) literature. Surprisingly, NIQE numbers are worse for target scene when compared to cut-and-paste (CP) or our reshaed cut-and-paste images.

Cut & Paste | Our | Cut & Paste | Our

Figure 10: **Failure Examples.** The wrongly shaded regions are marked in red. In the first example, DIP aggressively copies background shading onto the chair. However, the lego and the plant are reshaded accurately. In the second, the reason for failure is surface normals. The scene has two dominant normal directions – ground (pointing upwards) and the sky (pointed towards viewer). The lack of third direction forces DIP to copy shading either from the sky or the ground plane.

5 Discussion

Why could this work? It should be clear that corrections to inserted fragments are not veridical. As Liao et al. (2015; 2019) noted, corrected fragments often fool humans more effectively than physically accurate lighting, likely because humans attend to complex materials much more than to consistent illumination (Berzhanskaya et al., 2005). The alternative physics theory by Cavanagh & Alvarez (2005) argues that the brain employs a set of rules that are convenient, but not strictly physical, when interpreting a scene from retina image. When these rules are violated, a perception alarm is fired, or recognition is negatively affected (Gauthier et al., 1998). Otherwise, the scene "looks right". This means that humans may tolerate a considerable degree of estimation error, as long as it is of the right kind. By insisting that the image produces consistent inferences, we appear to be forcing errors to be "of the right kind".

Need For Speed. A desirable cut-and-paste reshading application would correct shading on inserted objects requiring no explicit 3D reasoning about either fragment or scene. It would do so in ways that produce consistent inferences of image analysis. Finally, it would be quick. We described a method that meets two of these three desiderata (DIP still requires minutes to render an image).

Casting Shadows. Our renderer can adjust the background with cast-shadows but often unrealistically. Therefore, we only show shading corrections on the fragment in this paper. Having better inferences (for eg., normal prediction network) should improve our results significantly. Also, better illumination modeling like Lighthouse (Srinivasan et al., 2020) and residual appearance networks (Sengupta et al., 2019) should also improve background shading changes. We also believe GAN based losses should help in correcting background shadows. This can also be combined with a few recent promising work in this direction (Izadinia & Seitz, 2020; Zhan et al., 2020).

Extending to Videos. Our method can reshade multiple frames at a time quite convincingly Fig 14 in Appendix. Extending to videos is a natural next step. Once the speed of our renderings is improved, we can also use our techniques for improving frame interpolation.

REFERENCES

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*, pp. 2317–2326, 2019.

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

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

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.

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

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*, pp. 8394–8403, 2020.

Paul Debevec. Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 1998.

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*, pp. 1301–1310, 2017.

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.

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.

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.

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.

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.

Hamid Izadinia and Steven M Seitz. Scene recomposition by learning-based icp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 930–939, 2020.

Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Josh Tenenbaum. Self-supervised intrinsic image decomposition. In *Advances in Neural Information Processing Systems*, 2017.

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.

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

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.

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

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

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

Edwin H Land. The retinex theory of color vision. *Scientific american*, 1977.

Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 1971.

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.

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.

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. 2020.

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*, pp. 3442–3449. IEEE, 2012.

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.

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.

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.

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.

Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. 2020.

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.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

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.

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

Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pp. 313–318. 2003.

Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 1985.

Vilayanur S Ramachandran. Perceiving shape from shading. *Scientific American*, 1988.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

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

Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8598–8607, 2019.

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.

Pratul P Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. 2020.

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.

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

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*, pp. 3789–3797, 2017.

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

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020.

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.

Fangneng Zhan, Shijian Lu, Changgong Zhang, Feiying Ma, and Xuansong Xie. Adversarial image composition with auxiliary illumination. *arXiv preprint arXiv:2009.08255*, 2020.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442*, 2016.

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.