

Neural Illumination: Lighting Prediction for Indoor Environments

Shuran Song Thomas Funkhouser
 Google and Princeton University

Abstract

This paper addresses the task of estimating the light arriving from all directions to a 3D point observed at a selected pixel in an RGB image. This task is challenging because it requires predicting a mapping from a partial scene observation by a camera to a complete illumination map for a selected position, which depends on the 3D location of the selection, the distribution of unobserved light sources, the occlusions caused by scene geometry, etc. Previous methods attempt to learn this complex mapping directly using a single black-box neural network, which often fails to estimate high-frequency lighting details for scenes with complicated 3D geometry. Instead, we propose “Neural Illumination,” a new approach that decomposes illumination prediction into several simpler differentiable sub-tasks: 1) geometry estimation, 2) scene completion, and 3) LDR-to-HDR estimation. The advantage of this approach is that the sub-tasks are relatively easy to learn and can be trained with direct supervision, while the whole pipeline is fully differentiable and can be fine-tuned with end-to-end supervision. Experiments show that our approach performs significantly better quantitatively and qualitatively than prior work.

1. Introduction

The goal of this paper is to estimate the illumination arriving at a location in an indoor scene based on a selected pixel in a single RGB image. As shown in Figure 1(a), the input is a low dynamic range RGB image and a selected 2D pixel, and the output is a high dynamic range RGB illumination map encoding the incident radiance arriving from every direction at the 3D location (“locale”) associated with the selected pixel (Figure 1(b)). This task is important for a range of applications in mixed reality and scene understanding. For example, the output illumination map can be used to light virtual objects placed at the locale so that they blend seamlessly into the real world imagery (Figure 8) and can assist estimating other scene properties, such as surface materials.

This goal is challenging because it requires a comprehensive understanding of the lighting environment. First,

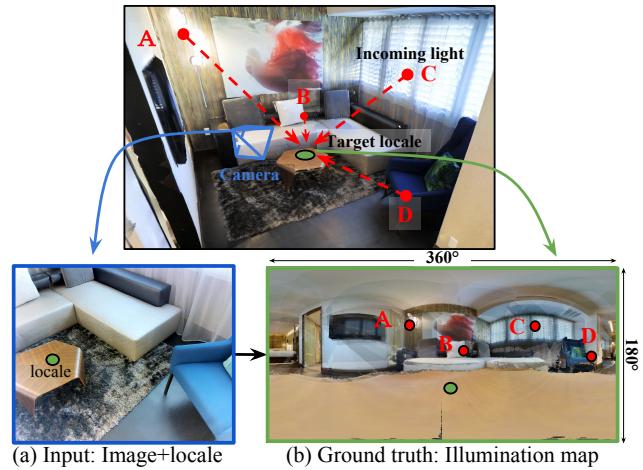


Figure 1. Given a single LDR image and a selected 2D pixel, the goal of **Neural Illumination** is to infer a panoramic HDR illumination map representing the light arriving from all directions at the locale. The illumination map is encoded as a spherical image parameterized horizontally by ϕ (0-360°) and vertically by θ (0-180°), where each pixel (e.g. A,B,C,D) stores the RGB intensity of light arriving at the “locale” from the direction (ϕ, θ) .

it requires understanding the 3D geometry of the scene in order to map between illumination observations at one 3D location (the camera) and another (the selected 3D locale). Second, it requires predicting the illumination coming from everywhere in the scene, even though only part of the scene is observed in the input image (e.g. the unobserved window in Figure 2). Third, it requires inferring HDR illumination from LDR observations so that virtual objects can be lit realistically. While it is possible to train a single neural network that directly models the illumination function end-to-end (from an input LDR image to an output HDR illumination map) [7], in practice optimizing a model for this complex function is challenging, and thus previous attempts have not been able to model high-frequency lighting details for scenes with complicated 3D geometry.

In this paper, we propose to address these challenges by decomposing the problem into three sub-tasks. First, to estimate the 3D geometric relationship between pixels in the input image and the output illumination map, we train a net-

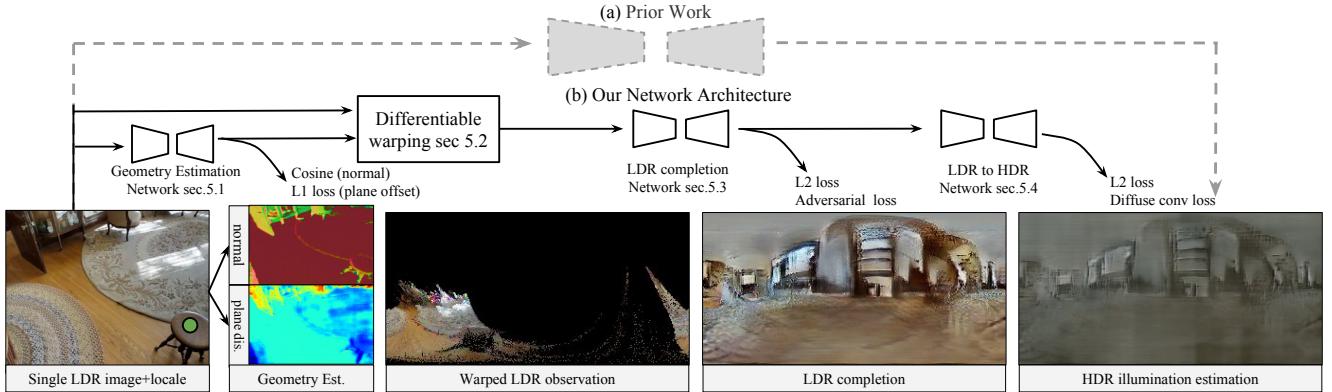


Figure 2. Neural Illumination. In contrast to prior work (a) [7] that directly trains a single network to learn the mapping from input images to output illumination maps, our network (b) decomposes the problem into three sub-modules: first the network takes a single LDR RGB image as input and estimate the 3D geometry of the observed scene. This geometry is used to warp pixels from the input image onto a spherical projection centered around an input locale. The warped image is then fed into LDR completion network to predict color information for the pixels in the unobserved regions. Finally, the completed image is passed through the LDR2HDR network to infer the HDR image. The entire network is differentiable and is trained with supervision end-to-end as well as for each intermediate sub-module.

work that estimates the 3D geometry from the observed image – the estimated geometry is then used to warp pixels from the input image to a spherical projection centered at the target locale to produce a partial LDR illumination map. Second, to estimate out-of-view and occluded lighting, we train a generative network that takes in the resulting partial illumination map and “completes” it – i.e., estimates the LDR illumination for all unobserved regions of the illumination map. Finally, to recover high dynamic range information, we train another network that maps estimated LDR colors to HDR light intensities. All these sub-modules are differentiable. They are first pre-trained individually with direct supervision and then fine-tuned end-to-end with the supervision of the final illumination estimation.

Our **key idea** is that by decomposing the problem into sub-tasks, it becomes practical to train an end-to-end neural network – each sub-module is able to focus on a relatively easier task and can be trained with direct supervision. The first sub-task is of particular importance – by predicting the 3D structure of the scene from the input image and using it to geometrically warp the input image such that it is spatially aligned with the output illumination map, we are able to enforce pixel-to-pixel spatial correspondence between the input and output representations, which has previously been shown to be crucial for other dense prediction tasks, such as image segmentation and edge detection.

To train and evaluate networks for this task, we have curated a benchmark dataset of paired input LDR images and output HDR illumination maps for a diverse set of locales in real-world scenes. In contrast to prior work, our dataset leverages panoramas captured densely in real-world scenes with HDR color and depth cameras [2]. We use the depth channel to warp and resample those panoramas at arbitrary locales to produce a set of 90,280 “ground truth” illumination maps observed in 129,600 images.

The primary contribution of our paper is introducing an end-to-end neural network architecture for illumination estimation (Neural Illumination) that decomposes the illumination estimation task into three sub-tasks. Our problem decomposition enables us 1) to provide both direct intermediate and end-to-end supervision, and 2) to convert the input observation into an intermediate representation that shares the pixel-wise spatial correspondence with the output representation. We show that this combination of neural network sub-modules leads to significantly better quantitative and qualitative results over prior work in experiments with our new benchmark dataset.

2. Related Work

Illumination estimation has been a long-standing problem in both computer vision and graphics. In this section, we briefly review work most relevant to this paper.

Capture-based Methods A direct way of obtaining the illumination of an environment is to capture the light intensity at a target location using a physical probe. Debevec *et al.* [3] first showed that photographs of a mirrored sphere with different exposures can be used to compute the illumination at the sphere’s location. Subsequent works show that beyond mirrored spheres, it is also possible to capture illumination using hybrid spheres [4], known 3D objects [24], object’s with know surface material [8], or even human faces [1] as proxies for light probes.

However, the process of physically capturing high-quality illumination maps can be expensive and difficult to scale, especially when the goal is to obtain training data for a dense set of visible locations in a large variety of environments. In this paper, we propose to use existing large-scale datasets with RGB-D and HDR panoramas (Matterport3D [2]) combined with image-based rendering methods

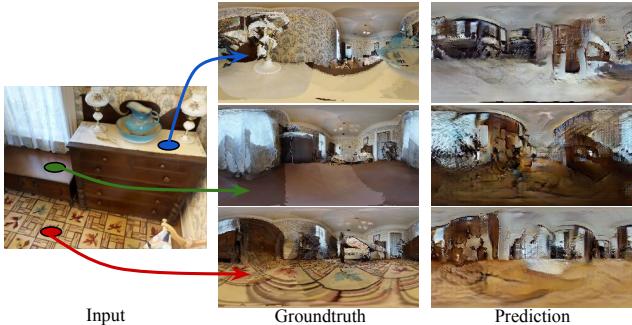


Figure 3. **Spatially varying illumination.** By using the 3D geometry, we can generate ground truth illumination for any target locale. As a result, our model is also able to infer spatially varying illumination conditioned on the target pixel location.

to generate a large training set of high-resolution illumination maps in diverse lighting environments.

Optimization-based Methods One standard approach to estimating illumination is to jointly optimize the geometry, reflectance properties, and lighting models of the scene in order to find the set of values that best explain the observed input image. However, directly optimizing all scene parameters is often a highly under-constrained problem – an error in one parameter estimation can easily propagate into another. Therefore to ease the optimization process, many prior methods either assume additional user-provided ground truth information as input or make strong assumptions about the lighting models. For example, Karsch *et al.* [12] uses user annotations for initial lighting and geometry estimates. Zhang *et al.* [26] uses manually-annotated light-source locations and assumes knowledge of depth information. Lombardi and Nishino [19] propose approximating illumination with a low-dimensional model, which subsequently has been shown to be sub-optimal for indoor scenes due to object reflective and geometric properties [7].

There are also works that explore the idea that similar images share similar illumination estimates. For example, Karsch *et al.* [13] uses image matching to find the most similar image crop from a panoramic database [25] and then use the lighting annotations on those panoramic images to predict out-of-view light sources. Khan *et al.* [14] directly flips observed HDR images to produce environment maps. In contrast, our system does not require additional user inputs or manual annotations, and it does not make any explicit assumptions about the scene content or its lighting models. Instead, we enable our learning-based model to learn illumination priors directly from data.

Learning-based Methods Deep learning has recently shown promising results on a number of computer vision tasks, including depth estimation [17, 15] and intrinsic image decomposition [27, 16]. Recently Gardner *et al.* [7] propose to formulate the illumination estimation function

as an end-to-end neural network. However, since the input and output representation of their network architecture does not share any immediate notion of pixel-to-pixel spatial correspondence, their model tends to generate illumination maps that reflect general color statistics of the training dataset, as opposed to important high-frequency lighting details. In contrast, our model predicts the 3D geometric structure of the scene and uses it to warp the observed image into an intermediate representation that encodes the input information in a way that is spatially aligned to the output illumination map. This results in the ability to fully utilize input information and preserve high-frequency lighting details. Moreover, Gardner *et al.*’s algorithm does not generate illumination conditioned on a selected target pixel (i.e., it produces only one solution for each input image). In contrast, our algorithm is able to recover the spatially varying illumination for any selected pixel (Figure 3).

Apart from differences in network design, Gardner *et al.* also suffers from the lack of accurate ground truth training data. Since their training data does not have depth, they use a sphere to approximate the scene geometry to warp a panorama to the target location. Moreover, since most of their training data is LDR, they use a binary light mask to approximate the bright HDR illumination during pre-training. While reasonable in the absence of 3D geometric and HDR information, these methods serve as weak approximations of ground truth. We address both of these data issues by directly training our model on a dataset that has both accurate 3D geometry and illumination information for a dense set of observations.

3. Problem formulation

We formulate illumination estimation as a pixel-wise regression problem modeled by a function $f: f(I|\ell) = H_\ell$ where I is an input LDR image of a scene, p is a selected pixel in the image. ℓ is the 3D location of the pixel, and H_ℓ is the output HDR illumination around ℓ . H_ℓ is represented as a spherical panoramic image with a 180° vertical FoV and 360° horizontal FoV. Each pixel $h(\phi, \theta) \in H_\ell$ of the panorama encodes the RGB intensity of incoming light to ℓ from the direction (ϕ, θ) . We model f as a feedforward convolutional neural network, the details of the network are described in Sec. 5. We train f on a large dataset of $\{I, \ell\}$ and H_ℓ^* pairs generated from Matterport3D (Sec. 4).

4. Generating a Dataset of Illumination Maps

Obtaining a large dataset of ground truth illumination maps for training is challenging. On the one hand, using physical probes to directly capture illumination at a target locale [3, 20, 4] provides accurate data, but scaling this capturing process across a diverse set of environments can be both costly and time-consuming. On the other hand, exist-

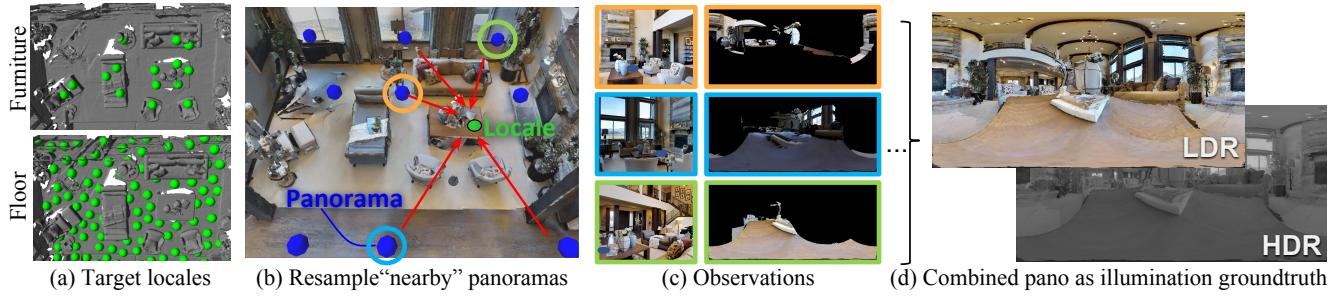


Figure 4. Ground truth illumination map generation. We generate reconstructions of over 90 different building-scale indoor scenes using HDR RGB-D images from the Matterport3D dataset [2]. From these reconstructions, we sample target locales (a) on supporting surfaces (floors and flat horizontal surfaces on furniture). For each locale, we use HDR and 3D geometric information from nearby RGB-D images to generate ground truth panoramic illumination maps.

ing panoramic datasets (*e.g.* [25]) provide a simple way to obtain illumination maps, but only at the camera locations around which the panoramas were captured.

Instead, we propose to leverage the HDR RGB-D images from the Matterport3D dataset [2] in combination with geometric warping to generate training data for arbitrary locales. Matterport3D contains 194,400 registered HDR RGB-D images arranged in 10,800 panoramas within 90 different building-scale indoor scenes. Since the panoramas provide ground truth HDR illumination maps for their center locations by direct observation, since they are acquired densely throughout each scene (separated by 2.5m or so), and since they have depth in addition to color, an RGB-D image-based rendering algorithm can reconstruct the illumination map *at any point in the scene* by warping and compositing nearby panoramas.

The first step of our dataset curation process is to sample a set of target locales. Ideally, the locales would cover the range of locations at which virtual objects could be placed in a scene. Accordingly, we densely sample locations 10cm above the surface of the input mesh and create a new locale if a) it is supported by a horizontal surface ($n_z > \cos(\pi/8)$), b) the support surface has semantic label $\in \{\text{floor, furniture}\}$, c) there is sufficient volumetric clearance to fit an object with radius of 10cm, d) it is not within 50cm of any previously created locale. For each locale, we backproject its location into every image I , check the depth channel to discard occlusions, and form a image-locale pair, $\{I, \ell\}$, for all others.

For each locale ℓ , we construct an illumination map H_ℓ^* using RGB-D image-based rendering. Though straightforward in principle, this process is complicated by missing depths at bright regions of the image (light sources, windows, strong specular highlights, etc.). A simple forward projection algorithm based on observed depths would omit these important elements of the illumination map. Therefore, we implemented a two-step process. During the first step, we estimate the distance to the closest surface in every direction $d(\phi, \theta)$ by forward mapping the depth channel of every input image I to ℓ , remembering the minimum dis-

tance in every direction, and filling holes with interpolation where no samples were mapped. Then, we reconstruct the illumination map for ℓ by resampling the HDR color channels of the input images via reverse mapping and blending the samples with weights proportional to $1/d^4$, where d is the distance between the camera and the locale. This process produces illumination maps with smooth blends of pixels from the nearest panoramas with holes filled by other panoramas further away. Overall, we generate 90,280 locales and 360,432 $\{I, \ell\}$ and H_ℓ^* pairs using this process. Figure 4 (a) shows examples for one scene.

Though the illumination maps produced this way are not always perfect (especially for highly specular surfaces), they have several favorable properties for training on our task. First, they are sampled from data collected by a large number of photographers [2] (mostly for real estate applications), and thus they contain a diverse set of lighting environments that would be difficult to gain access to otherwise. Second, they provide a unique illumination map for each 3D locale in a scene. Since multiple locales are usually visible in every single image, the dataset supports learning of spatial dependencies between pixel selections and illumination maps. For example, Figure 3 shows that our network is able to infer different illumination maps for different pixels selections in the same input image. Third, the “ground truth” illumination maps produced with our RGB-D warping procedure are more geometrically accurate than others produced with spherical warping [7]. As shown in Figure 5, our warping procedure is able to account for complex geometric structures and occlusions in the scene.

5. Network Architecture

In this section, we describe the convolutional neural network architecture used to model f , which consists of four sequential modules: 1) a geometry (RGB-to-3D) estimation module, 2) a differential warping module which warps the input RGB observation to the target locale using the estimated 3D information, 3) an out-of-view illumination estimation module, and 4) an LDR-to-HDR estimation module. Each module is pre-trained individually with its input and

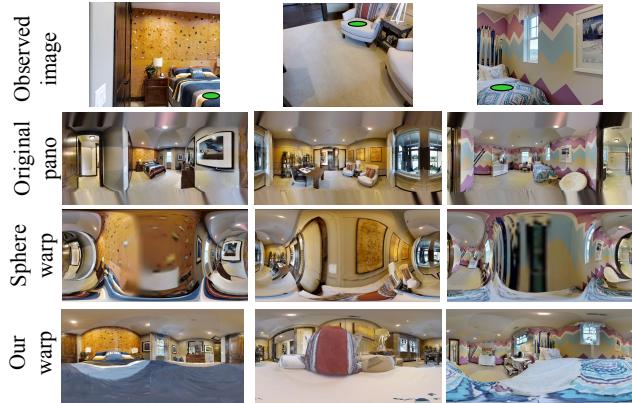


Figure 5. Comparison of warping methods. In our data generation process, we use 3D scene geometry to generate geometrically accurate ground truth illumination, which accounts for complex geometric structures and therefore more accurate than using spherical warping from 2D panoramas as in [7].

output pairs derived from ground truth information. Then all the modules are fine-tuned together end-to-end. Figure 2, shows the network architecture. By decomposing the network into sub-modules we allow each sub-module to focus on a relatively easier task with direct supervision. We find that providing both intermediate and end-to-end supervision is crucial for efficient learning.

5.1. Geometry Estimation

The geometry estimation module takes a single RGB image I as input and outputs a dense pixel-wise prediction of the visible 3D geometry G_I . Similar to Song *et al.* [22], G_I is represented with a “plane equation” for each pixel. Specifically, we feed I through a two-stream fully convolutional U-Net [21] to infer pixel-wise predictions of surface normals and plane offsets (*i.e.* distance-to-origin). We then pass both predicted outputs through a differentiable PN-layer [22] to convert the estimated surface normals and plane distances into a pixel-wise prediction of 3D locations. Direct supervision is provided to the 1) surface normal predictions via a cosine loss, 2) plane offset predictions via an ℓ_1 loss, and 3) final 3D point locations via an ℓ_1 to ensure consistency between the surface normal and plane offset predictions. Training labels are automatically obtained from the 3D data available in the Matterport3D dataset [2]. As shown in [22], this output representation provides strong regularization for large planar surfaces and is therefore able to produce higher quality predictions than directly predicting raw depth values [5, 15]. At the same time, it also maintains the flexibility of representing any surfaces – *i.e.*, is not limited to a fixed number of planar surfaces, as in [18].

5.2. Geometry-aware Warping

The next module uses the estimated scene geometry G_I to map the pixels in the input image I to a panoramic im-

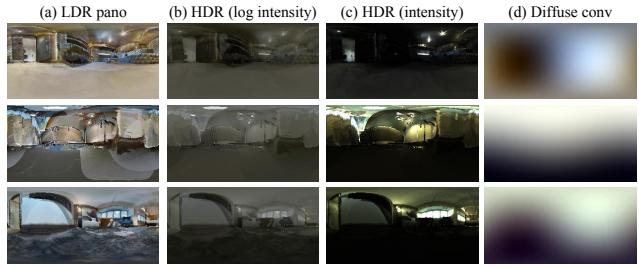


Figure 6. Examples of a) LDR image, b) log scaled HDR J , c) HDR intensity H , d) diffuse convolution of HDR intensity $D(H)$.

age ϕ_ℓ representing the unit sphere of rays arriving at ℓ . We do this warping through a forward projection using the estimated scene geometry and camera pose. The unit sphere projection that defines the panorama ϕ_ℓ is oriented upright along n_ℓ , which should be aligned with the gravity direction assuming that ℓ lays on a supporting surface (*e.g.* floors and flat horizontal surfaces on furniture). Image regions in ϕ_ℓ that do not have a projected pixel are set to -1. The resulting warped input observation is a panorama image with missing values that shares a pixel-wise spatial correspondence to the output illumination map. Since this warping module is entirely differentiable, we implement it as a single network layer.

5.3. LDR Panorama Completion

The third module takes the mapped observed pixels of ϕ_ℓ as input and outputs a dense pixel-wise prediction of illumination for the full panoramic image ψ_ℓ including both observed and unobserved pixels. ψ_ℓ is represented as a 3-channel LDR color panorama.

One of the biggest challenges for out-of-view illumination estimation comes from the multi-modal nature of the problem – there can be multiple possible solutions of ψ_ℓ with illumination patterns that result in similar observations. Therefore, in addition to providing only pixel-wise supervision, we train this module with adversarial loss using a discriminator network [9, 11]. This adversarial loss provides a learnable high-level objective by learning a loss that tries to classify if the output image is real or fake, while simultaneously training the generative model to minimize this loss. Our experiments show that this adversarial loss enables the network to produce and more realistic illumination outputs with sharper and richer details.

This module is implemented as a fully convolutional ResNet50 [10]. Since both the input and output of this module are represented as spherical panoramic images, we utilize distortion-aware convolutional filters that account for the different spherical distortion distributions for different regions of the image [23]. This distortion-aware convolution resamples the feature space according to the image distortion model in order to improve the translational invariance of the learned filters in the network.

5.4. LDR-to-HDR Estimation

The final module takes the predicted LDR illumination as input and outputs a dense pixel-wise prediction of HDR illumination intensities representing incident radiance in every direction at ℓ . This prediction is important because LDR images may have intensity clipping and/or tone-mapping, which would not be suitable for lighting virtual objects.

Like Eilertsen *et al.* [6], we formulate the LDR-to-HDR estimation as a pixel-wise regression problem, but instead of predicting the HDR value for only bright pixels and using a fixed function to map the rest of the pixels, our LDR-to-HDR module learns the mapping function for all pixels from the LDR space to the HDR space. The module is trained with supervision from: 1) a pixel-wise ℓ_2 loss L_{ℓ_2} , and 2) a diffuse convolutional loss L_d .

The pixel-wise ℓ_2 loss measures the visual error when re-lighting a perfectly *specular* surface at ℓ :

$$L_{\ell_2} = \frac{1}{N} \sum_{i=1}^N (J(i) - J^*(i))^2$$

where the J is log-scaled image of the final light intensity H , defined as:

$$H(i) = \begin{cases} J(i) * 65536 * 8e^{-8}, & J(i) \leq 3000 \\ 2.4e^{-4} * 1.0002^{(J(i)*65536-3000)}, & J(i) > 3000 \end{cases}$$

The diffuse convolutional loss measures the visual error when re-lighting a perfectly *diffuse* surface:

$$L_d = \frac{1}{N} \sum_{i=1}^N (D(H(i)) - D(H^*(i)))$$

where D is the diffuse convolution function defined as:

$$D(H, i) = \frac{1}{K_i} \sum_{\omega \in \Omega_i} H(\omega) s(\omega)(\omega \cdot \vec{n}_i)$$

and Ω_i is the hemisphere centered at pixel i on the illumination map, \vec{n}_i the unit normal at pixel i , and K_i the sum of solid angles on Ω_i . ω is a unit vector of direction on ω_i and $s(\omega)$ the solid angle for the pixel in the direction ω . This loss function is similar to the “cosine loss” function proposed by Gardner *et al.* [7], but rather than progressively increasing the Phong exponent value during training, we keep the Phong exponent value equal to 1. In our implementation, we reduce the memory usage by computing L_d on a downsized illumination map with average pooling.

The final loss is computed as $L = \lambda_1 L_{\ell_2} + \lambda_2 L_d$, where $\lambda_1 = 0.1$ and $\lambda_2 = 0.05$. By combining these two losses during training, we encourage our model to reproduce both low and high frequency illumination signals. Figure 6 shows examples of HDR images and their diffuse convolution.

6. Evaluation

We train and test our algorithm on the data generated from Section 4, using the train/test split provided by the

Matterport3D dataset [2]. The following experiments investigate qualitative and quantitative comparisons to prior work and results of ablation studies. More results and visualizations can be found in the supplementary material.

Evaluation metrics. We use the following evaluation metrics to quantitatively evaluate our predicted illumination maps H_ℓ :

- **Pixel-wise ℓ_2 distance error** is the sum of all pixel-wise ℓ_2 distances between the predicted H_ℓ and ground truth H_ℓ^* illumination maps. $\ell_2(\log)$ computes the ℓ_2 distance in the log intensity. Intuitively, this error measures the approximate visual differences observed when the maps are used to render a perfectly *specular* surface at the target locale.
- **Pixel-wise diffuse convolution error** is the sum of all pixel-wise ℓ_2 distances between $D(H_\ell)$ and $D(H_\ell^*)$. This error measures the approximate visual differences observed when the maps are used to render a perfectly *diffuse* surface at the target locale.

Comparisons to state-of-the-art. Table 1 shows quantitative comparisons of our approach to two alternative baselines: 1) Gardner *et al.* [7], and 2) a nearest neighbour retrieval method. Gardner *et al.* estimates the illumination condition of a given input image by training a single convolutional neural network with end-to-end supervision. We rotated each of the predicted panorama along the x-axis to align with ground truth coordinate frame before evaluation. Row 1 of Table 1 shows the performance of Gardner *et al.*’s model trained on their original LRD+HDR panorama dataset and tested on our test set. We also re-implement an image-to-image prediction network that is similar to Gardner *et al.*’s model and train it directly on our training data (LDR and full HDR illumination pairs) to remove potential dataset biases. This model [Im2Im network] achieves better performance than the original model but is still less accurate than ours. With a qualitative comparison (Figure 7,8), we can observe that by estimating and utilizing the 3D scene geometry, our algorithm is able to produce output illumination maps that contain much richer and more realistic high frequency details. Moreover, Gardner *et al.*’s algorithm does not allow users to input a specific target pixel – i.e., they generate only one lighting solution for each input image. In contrast, our algorithm is able to recover the spatially varying lighting distribution for any selected locale in the image, which can be quite different from one another (Figure 3).

Modularization v.s. additional supervision. While we show that our network is able to achieve better performance than the single end-to-end model, it is still unclear whether the performance gain comes from the additional supervision



Figure 7. **Qualitative Results** (Row 1) show the input image and selected locale. (Row 2,3) show the warped observation using ground truth an predicted geometry. (Row 4,5) show the completed LDR. (Row 6-10) show the final HDR illumination visualized with gamma correction ($\gamma=3.3$). We can observe that the illumination estimation from our approach is more accurate and also contain richer high frequency details.

Method	$\ell_2(\log)$	ℓ_2	diffuse
Gardner <i>et al.</i> [7]	0.375	0.977	1.706
Im2Im network	0.229	0.369	0.927
Nearest Neighbour	0.296	0.647	1.679
Ours	0.202	0.280	0.772

Table 1. Comparing the quantitative performance of our method to that of Gardner *et al.* [7] and a nearest neighbour retrieval method.

or the network modularization. To investigate this question, we trained an end-to-end network that takes in a single LDR image as input and directly outputs the completed 3D geometry, LDR images, and HDR images at the final layers. This network is trained with supervision for all three predictions but does not have any network decomposition. Table 2 shows the results. The performance gap between this network and ours demonstrates that naively adding all of the available supervision at the end of the network without proper network modularization and intermediate supervision does not work as well as our approach and generates

significantly lower-quality illumination maps.

	$\ell_2(\log)$	ℓ_2	diffuse
without	0.213	0.319	0.856
with (ours)	0.202	0.280	0.772

Table 2. Effects of modularization.

Comparisons to variants with oracles. To study how errors in intermediate predictions impacts our results, we execute a series of experiments where some data is provided by oracles rather than our predictions. In the first experiment, we trained a network that takes as input a LDR image already warped by a depth oracle and omits the first two modules (LDR+D→HDR). In a second experiment, we trained a version of that network that instead inputs a warped HDR image and omits execution of the last module. These networks utilize ground truth data, and thus are not fair comparisons. Yet, they provide valuable information about how well our network possibly could perform and which modules contribute most to the error. Looking

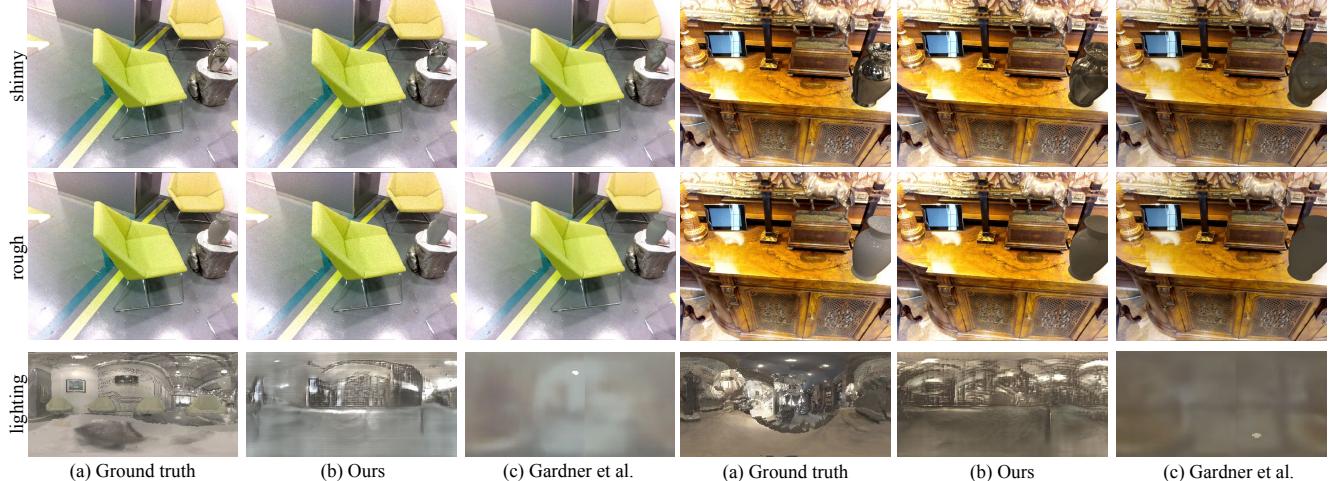


Figure 8. **Object relighting example.** Here we show qualitative comparisons of relighting results rendered by Mitsuba using the illumination maps from (a) ground truth, (b) our algorithm, and (c) Gardner *et al.* We show images rendered with two different surface materials composited over the original observations (the first and second rows) and the illumination maps (third row). Compared to (c), our algorithm is able to produce output illumination maps that contain much richer and more realistic high frequency detail. Of course, it also makes mistakes, for example by predicting extra light sources on the ceiling, which is incorrect but still plausible given the observation.

at Table 3, we see that providing ground truth depth improves our algorithm marginally (e.g., $\Delta\ell^2 = 0.011$), while also providing ground truth HDR improves it more (e.g., $\Delta\ell^2 = 0.043$). We conjecture it is because errors are concentrated on bright light sources. Overall, the performance of our algorithm is about halfway between the best oracled version [HDR+D→HDR] and the baselines in Table 1.

Method	$\ell^2(\log)$	ℓ^2	diffuse
LDR→HDR	0.202	0.280	0.772
LDR+D→HDR	0.188	0.269	0.761
HDR+D→HDR	0.131	0.212	0.619

Table 3. Comparisons to variants with oracles.

Effects of different losses. To study the effects of different loss functions, we evaluate the performance of the model “HDR+D→HDR” using different combinations of loss functions. Figure 4 shows quantitative and qualitative results. From the results we can observe that with only an ℓ^2 loss, the network tends to produce very blurry estimations that are close to the mean intensity of the input images. By adding the adversarial loss, the network starts to be able to infer more realistic high frequency signals and spotlights, but also introduces additional noises and errors in the prediction. By further adding a diffuse convolution loss [l2+gan+df], the network is able to predict overall more accurate illumination especially for the high intensity areas.

Conclusion and Future Work This paper presents “Neural Illumination,” an end-to-end framework for estimating high dynamic range illumination maps for a selected pixel in a low dynamic range image of an indoor scene. We propose to decompose the task into subtasks and train a net-

	l2 only	l2+gan	l2+gan+df	Ground truth
loss	$\ell^2(\log)$	ℓ^2	diffuse	
l2	0.116	0.235	0.691	
l2+gan	0.224	0.275	0.713	
l2+gan+df	0.131	0.212	0.619	

Table 4. Effects of different losses.

work module for: 1) inferring 3D scene geometry, 2) warping observations to illumination maps, 3) estimating unobserved illumination, and 4) mapping LDR to HDR. Experiments show that we can train a network with this decomposition that predicts illumination maps with better details and accuracy than alternative methods. While “Neural Illumination” is able to improve the accuracy of existing methods, it is still far from perfect. In particular, it often produces plausible illumination maps rather than accurate ones when no lights are observed directly in the input. Possible directions for future work include explicit modeling of surface material and reflective properties and exploring alternative 3D geometric representations that facilitate out-of-view illumination estimation through whole scene understanding.

Acknowledgments We thank Paul Debevec, John Flynn, Chloe LeGendre, and Wan-Chun Alex Ma for their valuable advice, Marc-Andr Gardner for producing results for comparison, Matterport for their dataset, and NSF 1539014/1539099 for funding.

References

- [1] D. A. Calian, J.-F. Lalonde, P. Gotardo, T. Simon, I. Matthews, and K. Mitchell. From faces to outdoor light probes. In *Computer Graphics Forum*, volume 37, pages 51–61. Wiley Online Library, 2018.
- [2] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *3DV*, 2017.
- [3] P. 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*, pages 189–198. ACM, 1998.
- [4] P. Debevec, P. Graham, J. Busch, and M. Bolas. A single-shot light probe. In *ACM SIGGRAPH 2012 Talks*, page 10. ACM, 2012.
- [5] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.
- [6] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM Transactions on Graphics (TOG)*, 36(6):178, 2017.
- [7] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambarotto, C. Gagné, and J.-F. Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (TOG)*, 36(6):176, 2017.
- [8] S. Georgoulis, K. Rematas, T. Ritschel, M. Fritz, T. Tuytelaars, and L. Van Gool. What is around the camera? In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [12] K. Karsch, V. Hedau, D. Forsyth, and D. Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 30(6):157, 2011.
- [13] K. Karsch, K. Sunkavalli, S. Hadap, N. Carr, H. Jin, R. Fonte, M. Sittig, and D. Forsyth. Automatic scene inference for 3d object compositing. *ACM Transactions on Graphics (TOG)*, 33(3):32, 2014.
- [14] E. A. Khan, E. Reinhard, R. W. Fleming, and H. H. Bültlhoff. Image-based material editing. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 654–663. ACM, 2006.
- [15] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.
- [16] Z. Li and N. Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–387, 2018.
- [17] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170, 2015.
- [18] M. Liu, X. He, and M. Salzmann. Geometry-aware deep network for single-image novel view synthesis. *arXiv preprint arXiv:1804.06008*, 2018.
- [19] S. Lombardi and K. Nishino. Reflectance and illumination recovery in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):129–141, 2016.
- [20] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski. *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010.
- [21] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [22] S. Song, A. Zeng, A. X. Chang, M. Savva, S. Savarese, and T. Funkhouser. Im2pano3d: Extrapolating 360 structure and semantics beyond the field of view. *Proceedings of 31th IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [23] K. Tateno, N. Navab, and F. Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 707–722, 2018.
- [24] H. Weber, D. Prévost, and J.-F. Lalonde. Learning to estimate indoor lighting from 3d objects. In *2018 International Conference on 3D Vision (3DV)*, pages 199–207. IEEE, 2018.
- [25] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba. Recognizing scene viewpoint using panoramic place representation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2695–2702. IEEE, 2012.
- [26] E. Zhang, M. F. Cohen, and B. Curless. Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics (TOG)*, 35(6):174, 2016.
- [27] T. Zhou, P. Krahenbuhl, and A. A. Efros. Learning data-driven reflectance priors for intrinsic image decomposition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3469–3477, 2015.