Fast Spatially-Varying Indoor Lighting Estimation

Mathieu Garon[⋆*], Kalyan Sunkavalli[†], Sunil Hadap[†], Nathan Carr[†], Jean-François Lalonde[⋆]
[⋆]Université Laval, [†]Adobe Research
mathieu.garon.2@ulaval.ca {sunkaval, hadap, ncarr}@adobe.com jflalonde@gel.ulaval.ca

1. Introduction

Estimating the illumination conditions of a scene is a challenging problem. An image is formed by conflating the effects of lighting with those of scene geometry, surface reflectance, and camera properties. Inverting this image formation process to recover lighting (or any of these other intrinsic properties) is severely underconstrained. Typical solutions to this problem rely on inserting an object (a light probe) with known geometry and/or reflectance properties in the scene (a shiny sphere [2], or 3D objects of known geometry [6, 18]). Unfortunately, having to insert a known object in the scene is limiting and thus not easily amenable to practical applications.

Previous work has tackled this problem by using additional information such as depth [1, 12], multiple images acquired by scanning a scene [7, 13, 20, 21] or user input [10]. However, such information is cumbersome to acquire. Recent work [5] has proposed a learning approach that bypasses the need for additional information by predicting lighting directly from a single image in an end-to-end manner. While [5] represents a practical improvement over previous approaches, we argue that this technique is still not amenable for use in more interactive scenarios, such as augmented reality (AR). First, it cannot be executed in real-time since it decodes full environment maps. Second, and perhaps more importantly, this approach produces a *single* lighting estimate for an image (more or less in the center of the image). However, indoor lighting is *spatially-varying*: light sources are in close proximity to the scene, thus creating significantly different lighting conditions across the scene due to occlusions and non-uniform light distributions.

In this work, we present a method that estimates spatially-varying lighting—represented as spherical harmonics (SH)—from a *single* image in real-time. Our method, based on deep learning, takes as input a single image and a 2D location in that image, and outputs the 5th-order SH coefficients for the lighting at that location. Our approach has three main advantages. First, spherical harmonics are a low-dimensional lighting representation (36 values for 5th-degree SH for each color channel), and can be predicted with a compact decoder architecture. Indeed, our experiments demonstrate that our network can predict 5th-degree SH coefficients in less than 20ms on a mobile GPU (Nvidia GTX970M). Second, the SH coefficients can directly be used by off-the-shelf shaders to achieve real-time relighting [14, 16]. Third, and perhaps more importantly, these local SH estimates directly embed local light visibility without the need for explicit geometry estimates. Our method therefore adapts to local occlusions and reflections without having to conduct an explicit reasoning on scene geometry. Note that while using SH constrains the *angular* frequency of the lighting we can represent, by having a different estimate for every scene location, our method does capture high-frequency *spatial* variations such as the shadowing under the desk in Figure 1(b).

Figure 1. Indoor lighting is spatially-varying. Methods that estimate global lighting [5] (left) do not account for local lighting effects resulting in inconsistent renders when lighting virtual objects. In contrast, our method (right) produces spatially-varying lighting from a single RGB image, resulting in much more realistic results.

To the best of our knowledge, our paper is the first to propose a practical approach for estimating spatially-varying lighting from a single indoor RGB image. Our approach enables a complete image-to-render augmented reality pipeline that automatically adapts to both local and global lighting changes at real-time framerates. In order to evaluate spatially-varying methods quantitatively, a novel, challenging dataset containing 79 ground truth HDR light probes in a variety of indoor scenes is made publicly available[1].

2. Dataset

In order to learn to estimate local lighting, we need a large database of images and their corresponding illumination conditions (light probes) measured at several locations

*Parts of this work were completed while Mathieu Garon was an intern at Adobe Research.

[1] https://lvsn.github.io/fastindoorlight/

in the scene. Relying on panorama datasets such as [5] unfortunately cannot be done since they do not capture local occlusions. While we provide a small dataset of real photographs for the evaluation of our approach (sec. 4.2), capturing enough such images to train a neural network would require a large amount of ressources. We therefore rely on realistic, synthetic data to train our neural network. In this section, we describe how we create our local light probe training data.

2.1. Rendering images

As in [22], we use the SUNCG [17] dataset for training. We do not use the Reinhard tonemapping algorithm [15] and instead use a simple gamma [11]. We now describe the corrections applied to the renders to improve their realism.

We render a total of 26,800 images, and use the same scenes and camera viewpoints as [22]. Care is taken to split the training/validation dataset according to houses (each house containing many rooms). Each image is rendered at 640×480 resolution using the Metropolis Light Transport (MLT) algorithm of Mitsuba [9], with 512 samples.

2.2. Rendering local light probes

For each image, we randomly sample 4 locations in the scene to render the local light probes. The image is split into 4 quadrants, and a random 2D coordinate is sampled uniformly in each quadrant (excluding a 5% border around the edges of the image). To determine the position of the virtual camera in order to render the light probe (the "probe camera"), a ray is cast from the scene camera to the image plane, and the first intersection point with geometry is kept. From that point, we move the virtual camera 10cm away from the surface, along the normal, and render the light probe at this location. Note that the probe camera axes are aligned with those of the scene camera—only a translation is applied.

3. Learning to estimate local indoor lighting

3.1. Main architecture for lighting estimation

We now describe our deep network architecture to learn spatially-varying lighting from an image. We require an input RGB image of 341×256 resolution and a specific coordinate in the image where the lighting is to be estimated. The image is provided to a "global" path in the CNN. A local patch of 150×150 resolution, centered on that location, is extracted and fed to a "local" path.

The global path processes the input image via the three first blocks of a pretrained DenseNet-121 network to generate a feature map. A binary coordinate mask, of spatial resolution 16×21, with the elements corresponding to the local patch set to 1 and 0 elsewhere, is concatenated as an additional channel to the feature map. The local path processes the local patch with a similar structure. Both global and local encoders share similar structures and use Fire modules [8].

The vectors coming from the global and local paths respectively are concatenated and processed by a fully-connected (FC) layer. The 5th-order SH coefficients in RGB are then predicted by another FC layer of dimensionality 36×3. We use an MSE loss on the SH coefficients.

SH Degree	Global (w/o mask)	Global (w mask)	Local	Local + Global (w mask)
0	0.698	0.563	0.553	**0.520**
1	0.451	0.384	0.412	**0.379**
2–5	0.182	**0.158**	0.165	0.159

Table 1. Ablation study on the network inputs. The mean absolute error (MAE) of each SH degree on the synthetic test set are reported.

SH Degree	$\mathcal{L}_{\text{i-sh}}$	$+\mathcal{L}_{\text{d-sh}}$	$+\mathcal{L}_{\text{rs-mse}}$ $+\mathcal{L}_{\text{rs-recons}}$	All
0	0.520	0.511	0.472	**0.449**
1	0.379	0.341	0.372	**0.336**
2–5	0.159	0.149	0.166	**0.146**
Degree 1 angle	0.604	0.582	0.641	**0.541**

Table 2. Comparing the mean absolute error (MAE) of the lighting SH degrees for each loss from 10,000 synthetic test probes.

3.2. Learning additional subtasks

It has recently been shown that similar tasks can benefit from joint training [19]. We now describe additional branches and losses that are added to the network to learn these related tasks, and in sec. 4.1 we present an ablation study to evaluate the impact of each of these subtasks.

Learning low-frequency probe depth Since lighting is affected by local visibility—for example, lighting under a table is darker because the table occludes overhead light sources—we ask the network to also predict SH coefficients for the low-frequency probe depth. To do so, we add another 36-dimensional output to the last FC layers. The loss for this branch is the MSE on the depth SH coefficients.

Learning patch albedo and shading To help disambiguate between reflectance and illumination, we also ask the network to decompose the local patch into its reflectance and shading intrinsic components. For this, we add a 3-layer decoder that takes in a $4 \times 4 \times 4$ vector from the last FC layer in the main branch, and reconstructs 7×7 pixel resolution (color) albedo and (grayscale) shading images.

Adapting to real data We apply unsupervised domain adaptation [4] to adapt the model trained on synthetic SUNCG images to real photographs.

4. Experimental validation

We now present an extensive evaluation of our network design as well as qualitative and quantitative results on a

		All	Center	Off-center
RMSE	global-[5]	0.081 ± 0.015	0.079 ± 0.021	0.086 ± 0.019
	local-[5]	0.072 ± 0.013	0.086 ± 0.027	0.068 ± 0.019
	Ours	**0.049 ± 0.006**	**0.049 ± 0.019**	**0.051 ± 0.012**
siRMSE	global-[5]	0.120 ± 0.013	0.124 ± 0.018	0.120 ± 0.031
	local-[5]	0.092 ± 0.017	0.120 ± 0.035	0.084 ± 0.016
	Ours	**0.062 ± 0.005**	**0.072 ± 0.011**	**0.055 ± 0.009**

Table 3. Comparing the relighting error between each method.

new benchmark test set. We evaluate our system's accuracy at estimating 5th order SH coefficients. We chose order 5 after experimenting with orders ranging from 3 to 8, and empirically confirming that order 5 SH lighting gave us a practical trade-off between rendering time and visual quality (including shading and shadow softness). In principle, our network can be easily extended to infer higher order coefficients.

4.1. Validation on synthetic data

A non-overlapping test set of 9,900 probes from 2,800 synthetic images (sec. 2) rendered from different houses is used to perform two ablation studies to validate the design choices in the network architecture (sec. 3.1) and additional subtasks (sec. 3.2).

First, we evaluate the impact of having both global and local paths in the network, and report the mean absolute error (MAE) in SH coefficient estimation in tab. 1. For this experiment, the baseline ("Global (w/o mask)") is a network that receives only the full image, similar to Gardner et al. [5]. Without local information, the network predicts the average light condition of the scene and fails to predict local changes, thus resulting in low accuracy. Lower error is obtained by concatenating the coordinate mask to the global DenseNet feature map ("Global (w mask)").

Second, tab. 2 shows that learning subtasks improves the performance for the light estimation task [19]. Activating the MSE loss on the low frequency probe depth significantly improves the directional components of the SH coefficients, but has little impact on the degree 0. Conversely, training with an albedo/shading decomposition task improves the ambient light estimation (SH degree 0), but leaves the directional components mostly unchanged.

4.2. A dataset of real images and local light probes

To validate our approach, we captured a novel dataset of real indoor scenes and corresponding, spatially-varying light probes. For each scene, an average of 4 HDR light probes are subsequently captured by placing a 3-inch diameter chrome ball [3] at different locations, and shooting the entire scene with the ball in HDR once more. In all, a total of 20 indoor scenes and 79 HDR light probes were shot. In the following, we use the dataset to compare our methods quantitatively, and through a perceptual study.

4.3. Comparison on real photographs

We use the real dataset from sec. 4.2 to compare our method against two versions of the approach of Gardner et al. [5], named *global* and *local*. The global version is their original algorithm, which receives the full image as input and outputs a single, global lighting estimate. For a perhaps fairer comparison, we make their approach more local by giving it as input a crop containing a third of the image with the probe position as close as possible to the center. We also show qualitative comparison against Barron and Malik [1]. While their approach yields spatially-varying SH lighting, it typically produces conservative estimates that do not capture the spatial variation in lighting accurately. In contrast, our method requires only RGB input, runs in real-time, and yields more realistic lighting estimates. The results on NYU-v2 are presented in the supplementary material.

Relighting error We compare all methods by rendering a diffuse bunny model with the ground truth environment map, with the algorithms outputs, and compute error metrics on the renders. A comparison against [5] is provided in tab. 3. To provide more insight, we further split the light probes in the dataset into two different categories: the *center* and *off-center* probes. The center probes were determined, by manual inspection, to be those close to the center of the image, and not affected by the local geometry or close light sources. Our method outperforms both versions of [5].

User study We further conduct a user study to evaluate whether the quantitative results obtained in the previous section are corroborated perceptually. For each of the 3 techniques (ours, global-[5], local-[5]), we show the users pairs of images: the reference image rendered with the ground truth light probe, and the result rendered with one of the lighting estimates. Each user is presented with all of the 20 scenes, and for each scene a random probe and a random technique is selected. The study was conducted using Amazon Mechanical Turk. Our method achieves a confusion of 35.8% (over a maximum of 50%), compared to 28% and 31% for the local and global versions of [5]. Our method outperforms both method on the probes affected by local geometry only. It achieves 34.5% compared to 27.1% and 29.5% on the two versions of [5].

5. Conclusion and Future Work

We present a real-time method, particularly suitable for AR, to predict local lighting for indoor scenes. As demonstrated via extensive evaluations on synthetic and real data, our method significantly outperforms previous work. A future direction will be to explore different lighting representations to improve the angular frequency of our predictions leading to crisper shadows, and ultimately suitable reflection maps, for a seamless physically-based rendering pipeline and AR experience.

References

[1] J. Barron and J. Malik. Intrinsic scene properties from a single rgb-d image. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[2] 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*, ACM Transactions on Graphics (SIGGRAPH), pages 189–198, 1998.

[3] P. E. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *ACM Transactions on Graphics (SIGGRAPH)*, page 31. ACM, 2008.

[4] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, 2015.

[5] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 9(4), 2017.

[6] S. Georgoulis, K. Rematas, T. Ritschel, M. Fritz, T. Tuytelaars, and L. Van Gool. What is around the camera? In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[7] L. Gruber, T. Richter-Trummer, and D. Schmalstieg. Real-time photometric registration from arbitrary geometry. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2012.

[8] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[9] W. Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org.

[10] K. Karsch, V. Hedau, D. Forsyth, and D. Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (SIGGRAPH asia)*, 30(6):1, 2011.

[11] Z. Li and N. Snavely. CGIntrinsics: Better intrinsic image decomposition through physically-based rendering. In *European Conference on Computer Vision (ECCV)*, 2018.

[12] R. Maier, K. Kim, D. Cremers, J. Kautz, and M. Nießner. Intrinsic3d: High-quality 3d reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3114–3122, 2017.

[13] R. Monroy, M. Hudon, and A. Smolic. Dynamic environment mapping for augmented reality applications on mobile devices. In F. Beck, C. Dachsbacher, and F. Sadlo, editors, *Vision, Modeling and Visualization*. The Eurographics Association, 2018.

[14] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *ACM Transactions on Graphics (SIGGRAPH)*, pages 497–500. ACM, 2001.

[15] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda. Photographic tone reproduction for digital images. *ACM transactions on graphics (SIGGRAPH*, 21(3):267–276, 2002.

[16] P.-P. Sloan, J. Kautz, and J. Snyder. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. In *ACM Transactions on Graphics (SIGGRAPH)*, volume 21, pages 527–536. ACM, 2002.

[17] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] H. Weber, D. Prévost, and J.-F. Lalonde. Learning to estimate indoor lighting from 3D objects. In *International Conference on 3D Vision (3DV)*, 2018.

[19] A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] E. Zhang, M. F. Cohen, and B. Curless. Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics (SIGGRAPH asia)*, 35(6), 2016.

[21] E. Zhang, M. F. Cohen, and B. Curless. Discovering point lights with intensity distance fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[22] Y. Zhang, S. Song, E. Yumer, M. Savva, J.-Y. Lee, H. Jin, and T. Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.