DeepLight: Learning Illumination for Unconstrained Mobile Mixed Reality

Chloe LeGendre
Google Inc.
chlobot@google.com

Wan-Chun Ma
Google Inc.
nitroboy@google.com

Graham Fyffe
Google Inc.
fyffe@google.com

John Flynn
Google Inc.
jflynn@google.com

Laurent Charbonnel
Google Inc.
lcharbonnel@google.com

Jay Busch
Google Inc.
jbusch@google.com

Paul Debevec
Google Inc.
debevec@google.com

(a) training data (b) input image (c) output lighting (d) rendered object (e) real object (f) rendered object (g) real object

Figure 1: Given an arbitrary low dynamic range (LDR) input image captured with a mobile device (b), our method produces omnidirectional high dynamic range lighting (c, lower) useful for rendering and compositing virtual objects into the scene. We train a CNN with LDR images (a) containing three reflective spheres, each revealing different lighting cues in a single exposure. (d) and (f) show renderings produced using our lighting, closely matching photographs of real 3D printed and painted objects in the same scene (e, g).

ABSTRACT

We present a learning-based method to infer plausible high dynamic range (HDR), omnidirectional illumination given an unconstrained, low dynamic range (LDR) image from a mobile phone camera with a limited field of view (FOV). For training data, we collect videos of various reflective spheres placed within the camera's FOV, leaving most of the background unoccluded, leveraging that materials with diverse reflectance functions reveal different lighting cues in a single exposure. We train a deep neural network to regress from the LDR background image to HDR lighting by matching the LDR ground truth sphere images to those rendered with the predicted illumination using image-based relighting, which is differentiable. Our inference runs at interactive frame rates on a mobile device, enabling realistic rendering of virtual objects into real scenes for mobile mixed reality. Training on auto-exposed and white-balanced videos, we improve the realism of rendered objects compared to the state-of-the art methods for both indoor and outdoor scenes.

SIGGRAPH '19 Talks, July 28 - August 01, 2019, Los Angeles, CA, USA

ACM ISBN 978-1-4503-6317-4/19/07.
https://doi.org/10.1145/3306307.3328173

CCS CONCEPTS

• **Computing methodologies** → **Image-based rendering**; **Scene understanding**;

KEYWORDS

Image-based lighting, augmented reality, lighting estimation

ACM Reference Format:
Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. 2019. DeepLight: Learning Illumination for Unconstrained Mobile Mixed Reality. In *Proceedings of SIGGRAPH '19 Talks*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3306307.3328173

1 INTRODUCTION AND RELATED WORK

Compositing rendered virtual objects into photographs or videos is a fundamental technique in mixed reality and film production. The realism of a composite depends on both geometric factors and lighting. An object "floating in space" rather than placed on a surface will immediately appear fake; similarly, a rendered object that is too bright, too dark, or lit from a direction inconsistent with other objects in the scene can be just as unconvincing. In this work, we propose a method to estimate plausible illumination from mobile phone images or video to convincingly light synthetic 3D objects for real-time compositing.

Estimating scene illumination from a single photograph with low dynamic range (LDR) and a limited field of view (FOV) is a challenging, under-constrained problem. One reason is that an object's appearance in an image is the result of the light arriving from the full sphere of directions around the object, including from those outside the camera's FOV. However, in a typical mobile phone video, only 6% of the panoramic scene is observed by the camera. Furthermore, even light sources within the FOV will likely be too bright to be measured in a single exposure if the rest of the scene is well-exposed, saturating the image sensor due to limited dynamic range and thus yielding an incomplete record of relative scene radiance. To measure this missing information, Debevec [1998] merged omnidirectional photographs captured with different exposure times and lit synthetic objects with these high dynamic range (HDR) panoramas using global illumination rendering. But in the absence of such measurements, professional lighting artists often create convincing illumination by reasoning on cues like shading, geometry, and context, suggesting that a background image alone may provide sufficient information for plausible lighting estimation.

As with other challenging visual reasoning tasks, convolutional neural networks (CNNs) comprise the state-of-the-art techniques for lighting estimation from a limited-FOV, LDR image, for both indoor [Gardner et al. 2017] and outdoor [Hold-Geoffroy et al. 2017] scenes. Naïvely, many pairs of background images and lighting (HDR panoramas) would be required for training; however, capturing HDR panoramas is complex and time-consuming, so no such dataset exists for both scene types. For indoor scenes, Gardner et al. [2017] first trained a network with many LDR panoramas, and then fine-tuned it with 2100 captured HDR panoramas. For outdoor scenes, Hold-Geoffroy et al. [2017] fit a sky model to LDR panoramas for training data. We also use a CNN, but our model generalizes to both indoor and outdoor scenes, requires no HDR imagery, and runs at 12-20 fps on a mobile device.

2 METHOD

We capture training data as LDR images with three spheres held within the bottom portion of the camera's FOV (Fig. 2), each with a different material that reveals different cues about the scene's ground truth illumination. For instance, a mirrored sphere reflects omnidirectional, high-frequency lighting, but, in a single exposure, bright light source reflections usually saturate the sensor so their intensity and color are misrepresented. A diffuse gray sphere, in contrast, reflects blurred, low-frequency lighting, but captures a relatively complete record of the total light in the scene and its general directionality. We regress from the portion of the image unoccluded by the spheres to the HDR lighting, training our network by minimizing the difference between the LDR ground truth sphere images and their appearances *rendered* with the estimated lighting. We first measure each sphere's reflectance field [Debevec et al. 2000]. Then, during training, we render the spheres with the estimated HDR lighting using image-based relighting (IBRL) [Debevec et al. 2000], which is differentiable. Furthermore, we add an adversarial loss term to improve recovery of plausible high-frequency illumination. As only one exposure comprises each training example, we can capture *videos*, increasing the volume of real-world training data and giving a prior on the camera's auto-exposure and white-balance. For further details of the network architecture and training, please see the supplemental materials.

Figure 2: Left: Capture apparatus. Center: Example frame. Right: Processed data (top: input; bottom: ground truth).

3 RESULTS

Ground truth comparisons: In Fig. 3, we show ground truth spheres compared with those rendered using IBRL and our HDR lighting inference, for each BRDF (**d**iffuse, matte **s**ilver, **m**irror), corresponding to the 25th, 50th, and 75th percentiles for rendering reconstruction loss, for unseen indoor (UI) and outdoor (UO) scenes. In Fig. 1, we compare off-line renders using our HDR lighting inference with the appearance of a real object in the same scene.

Additional results: For quantitative evaluation, additional renderings, comparisons with previous work, a lighting inference mobile demo, and limitations, please see the supplemental materials.

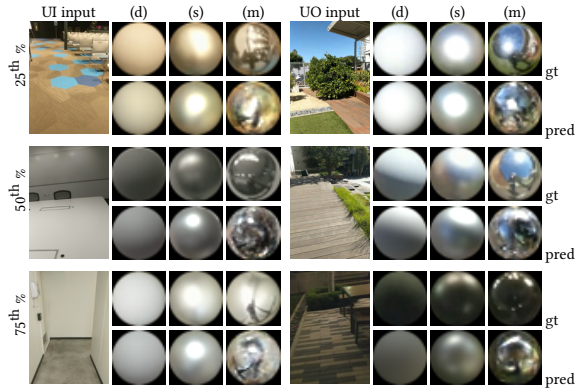

Figure 3: Qualitative comparisons between ground truth spheres and IBRL renderings using our CNN-based HDR lighting inference. Examples for 25th, 50th, and 75th percentiles for rendering loss.

REFERENCES

Paul Debevec. 1998. 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, 189–198.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 145–156.

Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to Predict Indoor Illumination from a Single Image. *ACM Trans. Graph.* 36, 6, Article 176 (Nov. 2017), 14 pages. https://doi.org/10.1145/3130800.3130891

Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. 2017. Deep outdoor illumination estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, Vol. 2.