DeepLight: Learning Illumination for Unconstrained Mobile Mixed Reality

Chloe LeGendre[1,2 *] Wan-Chun Ma[1] Graham Fyffe[1] John Flynn[1]
Laurent Charbonnel[1] Jay Busch[1] Paul Debevec[1]
[1]Google Inc. [2]USC Institute for Creative Technologies

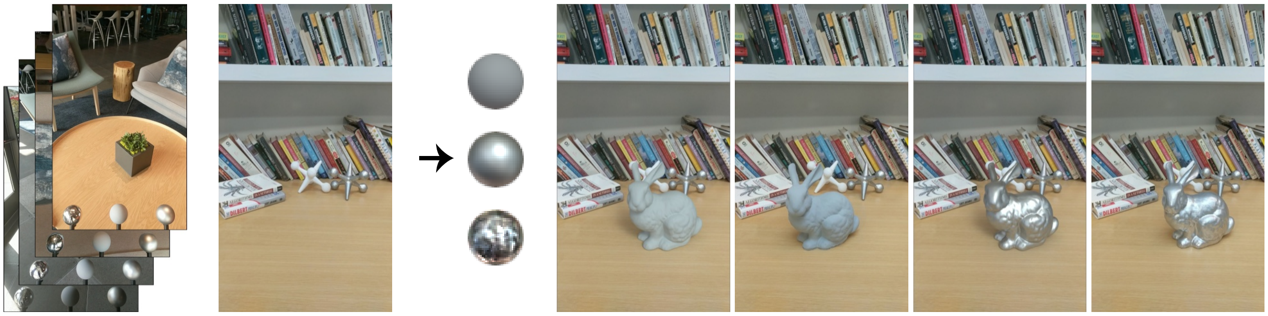

(a) training data (b) input image (c) output lighting (d) rendered object (e) real object (f) rendered object (g) real object

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 automatically 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.

1. Introduction

Compositing rendered virtual objects into photographs or videos is a fundamental technique in mixed reality, visual effects, and film production. The realism of a composite depends on both geometric and lighting related factors. 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 directions outside the camera's FOV. However, in a typical mobile phone video, only 6% of the panoramic scene is observed by the camera (see Fig. 2). Furthermore, even light sources within the FOV will likely be too bright to be measured properly 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 [10] 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 illumina-

[*] Work completed while interning at Google.

tion by reasoning on cues like shading, geometry, and context, suggesting that a background image alone may provide sufficient information for plausible lighting estimation.

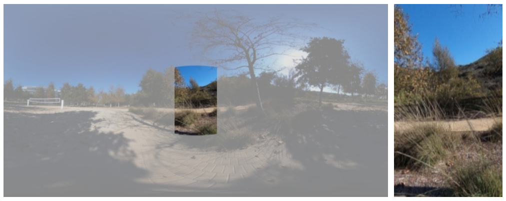


Figure 2: The field of view (FOV) of mobile phone video (inset shown in full color), relative to the 360° environment.

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 [14] and outdoor [18] 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. [14] first trained a network with many LDR panoramas [55], and then fine-tuned it with 2100 captured HDR panoramas. For outdoor scenes, Hold-Geoffroy et al. [18] 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 and requires no HDR imagery.

In this work, our training data is captured as LDR images with three spheres held within the bottom portion of the camera's FOV (Fig. 3), 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 the 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 as in [12]. Then, during training, we render the spheres with the estimated HDR lighting using image-based relighting [12, 35], 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* of real-world scenes, which increases the volume of training data and gives a prior on the automatic exposure and white-balance of the camera.

For a public benchmark, we collect 200k new images in indoor and outdoor scenes, each containing the three different reflective spheres. We show on a random subset that our method out-performs the state-of-the-art lighting estimation techniques for both indoor and outdoor scenes for mobile phone imagery, as our inferred lighting more accurately renders synthetic objects. Furthermore, our network runs at interactive frames rates on a mobile device, and, when used in combination with real-time rendering techniques, enables more realistic mobile mixed reality composites.

In summary, our key contributions are:

- A data collection technique and dataset of paired lighting reference spheres and background images (200k examples) for training a lighting estimation algorithm.
- A CNN-based method to predict plausible omnidirectional HDR illumination from a single unconstrained image. To the best of our knowledge, ours is the first to generalize to both indoor and outdoor scenes.
- A novel image-based relighting rendering loss function, used for training the HDR lighting inference network using *only* LDR data.

2. Related work

Debevec [10] rendered synthetic objects into photographs of real-world scenes using HDR panoramas as lighting. These can be captured by photographing a mirrored sphere or stitching together wide-angle views using multiple exposures [10, 46]. Recording HDR video of a mirror ball [52, 51] has been used for real-time capture of image-based lighting environments. Our goal is to estimate HDR lighting given only a single LDR image with a limited, but fixed, FOV. Key to our technique is that spheres with diverse reflectance functions (BRDFs) reveal different lighting cues, enabling us to record training data using a standard LDR video stream. This has been previously leveraged for sun intensity recovery from clipped panoramas using a diffuse, gray sphere [39, 11].

The appearance of a scene depends on its geometry, reflectance properties, and lighting, as well as the camera's exposure, color balance, and depth-of-field. The joint recovery of geometry, reflectance, and lighting, known as the inverse rendering problem, has been a core computer vision challenge [59, 38]. Intrinsic image decomposition [4] separates an image into shading and reflection; however, shading is an *effect* of lighting, not its direct observation. While recent approaches jointly inferred material reflectance and illumination from an object comprised of an unknown material [34, 29], one or more images of a segmented object [53, 31], specular objects of a known class [40, 15], or with measured or known geometry [30, 54, 33, 17, 3], we estimate lighting from unconstrained images with unknown geometry and arbitrarily complex scenes.

Khan et al. [24] projected a limited-FOV HDR image onto a hemisphere and flipped it to infer 360° lighting. For

LDR images, Karsch et al. [23] estimated a scene's geometry and diffuse albedo, detected in-view light sources, and, for unseen lights, found a matching LDR panorama from a database [55]. They promoted the result to HDR, minimizing a diffuse scene rendering loss. For indoor scenes, Gardner et al. [14] learned a mapping from a limited FOV LDR image to HDR lighting using a CNN. Noting the lack of HDR panoramas, they leveraged the same LDR panorama dataset [55] to regress first from the input image to a LDR panorama and light source locations and then refined the model for light source intensities with 2100 new, captured HDR panoramas. Though demonstrating state-of-the-art results, they noted two key limitations. First, the predicted LDR panorama and HDR light sources were white-balanced to match the input image using the Gray World assumption [6]. Second, renderings improved when an artist manually tuned the predicted lighting intensity. We propose a novel rendering-based loss function that allows our network to learn both the colors and intensities of the incident illumination relative to the input image, without HDR imagery. Furthermore, we propose a lighting model that generalizes to both indoor and outdoor scenes, though outdoor HDR lighting estimation from a single image or from a LDR panorama has also received attention, as the sun and sky afford lower dimensional lighting parameterizations [26, 27, 18, 60]. Cheng et al. [8] estimated lighting from opposing views within a panorama for indoor and outdoor scenes, but did not consider single image inputs.

Several recent works estimate lighting from faces, modeling image formation via rendering within the "decoder" of an encoder-decoder architecture [7, 49, 48, 61, 42, 43]. However, all have relied on simple or low frequency shading models. In contrast, we render objects during training using image-based relighting (IBRL) [12, 35], forming new images as a linear combination of reflectance basis images, avoiding an analytic shading model altogether. Xu et al. [56] trained a network to perform IBRL, jointly *learning* a low-dimensional reflectance basis and renderer, rather than applying IBRL as a fixed function as we do. Hold-Geoffroy et al. [18] and Cheng et al. [8] used a synthetic Lambertian reflectance basis in a rendering loss term but did not use a photographed basis or consider multiple BRDFs.

3. Method

Here we describe how we acquire our training data, our network architecture, and the loss functions of our end-to-end lighting estimation method.

3.1. Training Data Acquisition and Processing

Gardner et al. [14] fine-tuned a pre-trained network using 2100 HDR panoramas, fewer examples than would be typically required for deep learning without pre-training. However, our key insight is that we can infer HDR lighting from *only* LDR images with reference objects in the scene, provided they span a range of BRDFs that reveal different lighting cues. Thus, we collect LDR images of indoor and outdoor scenes, where each contains three spheres located in the bottom portion of the camera's FOV, occluding as little of the background as possible (Fig. 3, center). The three spheres are plastic holiday ornaments with diverse finishes that differently modulate the incident illumination: mirrored silver, matte silver (rough specular), and diffuse gray (spray-painted), measured as 82.7%, 64.4%, and 34.5% reflective respectively. We built a capture rig to fix the sphere-to-phone distance, stabilizing the sphere positions in each image (see Fig. 3, left).

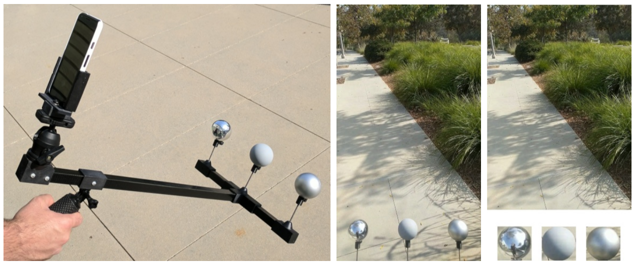


Figure 3: Left: Capture apparatus. Center: Example frame. Right: Processed data (top: input; bottom: ground truth).

As we require only LDR input imagery, we collect portrait HD (1080 × 1920) *video* at 30 fps, rather than static photographs. This increases the speed of training data acquisition compared with HDR panoramic photography, enabling the capture of millions of images, albeit with significant redundancy for adjacent frames. The videos feature automatic exposure and white balance, providing a prior to help disambiguate color, reflectance, and illumination.

We locate the three spheres in each video frame by detecting circular boundaries in the optical flow field between neighboring frames (see supplemental materials for more details), though marker-based tracking could also be used. We re-sample cropped images of the spheres using an idealized camera model oriented towards the sphere center with a view frustum tangent to the sphere on all four sides to eliminate perspective distortion. For the background images, we remove the lower 20% of each frame during both training and inference. The final training data consists of cropped background images, each paired with a set of three cropped spheres, one per BRDF (Fig. 3, right).

3.2. Network Architecture

The input to the model is an unconstrained LDR, gamma-encoded image captured with a mobile phone, resized from the native cropped resolution of 1080×1536 to 135×192 and normalized to the range of $[-0.5, 0.5]$. Our architecture is an encoder-decoder type, where the encoder includes fast depthwise-separable convolutions [20]. We use the first 17 MobileNetV2 [41] layers, processing

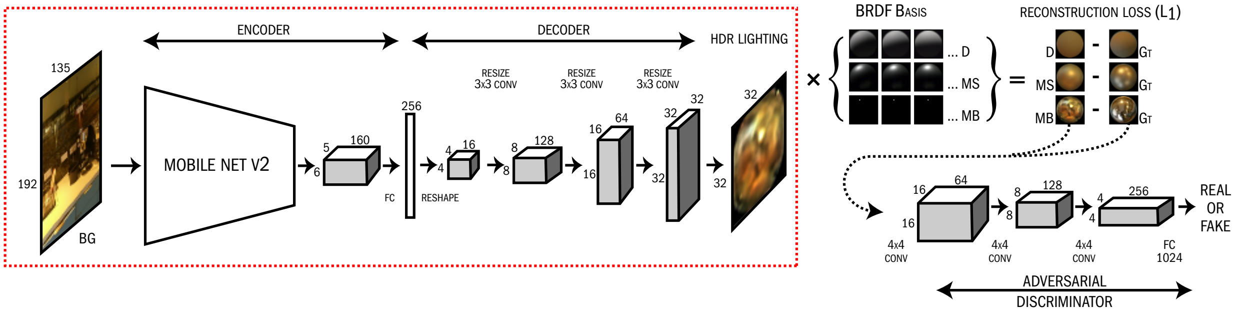


Figure 4: Overview of our network. We regress to HDR lighting from an LDR, limited-FOV input image captured with a mobile device. We include a multi-BRDF image-based relighting reconstruction loss for a diffuse(D), matte silver(MS), and mirror ball(MB) and an adversarial loss for the mirror ball only. Only the part outlined in red occurs at inference time.

the output feature maps with a fully-connected (FC) layer to generate a latent vector of size 256. For the decoder, we reshape this vector and upsample thrice by a factor of two to generate a 32×32 color image of HDR lighting. We regress to natural log space illumination as the sun can be more than five orders of magnitude brighter than the sky [46]. Although we experimented with fractionally-strided convolutions, bilinear upsampling with convolutions empirically improved our results. We train the network to produce omnidirectional lighting in the mirror ball mapping [39], where each pixel in image space represents an equal solid angular portion of a sphere for direction (θ, ϕ). Thus, the corners of the output image are unused, but this mapping allows for equal consideration of all lighting directions in the loss function, if desired. For network details, see Fig. 4.

3.3. Reflectance Field Acquisition

Debevec et al. [12] introduced the 4D reflectance field $R(\theta, \phi, x, y)$ to denote the image of a subject with pixels (x, y) as lit from any lighting direction (θ, ϕ) and showed that taking the dot product of the reflectance field with an HDR illumination map relights the subject to appear as they would in that lighting. During training, we use this method to render spheres with the predicted HDR lighting. We photograph reflectance fields for the matte silver and diffuse gray spheres using a computer-controllable sphere of white LEDs [32], spaced 12° apart at the equator. This produces an anti-aliased reflectance field for the diffuse and matte silver sphere; however, this LED spacing aliases the mirror BRDF. As we infer lighting in a mirror ball mapping, we instead construct the mirror ball basis as a set of 32×32 one-hot matrices of size 32×32, scaled by its measured reflectivity. We convert the lighting bases for the other BRDFs to the same geometric and relative radiometric space. The photographed bases are normalized based on the incident light source color and converted to the mirror ball mapping, accumulating energy from the photographs for each new lighting direction i for the set of directions on the 32×32 mirror ball using a Phong lobe ($n = 64$) and super-sampling with a 4×4 grid of directions on a sphere.

3.4. Loss Function

To train the lighting prediction network, we minimize an image-based relighting loss and add an adversarial loss to ensure inference of plausible high-frequency illumination.

Image-based relighting rendering loss: We train the network by minimizing the reconstruction loss between the ground truth sphere images I and rendered spheres lit with the predicted HDR lighting. With the reflectance fields $R(\theta, \phi, x, y)$, pixel values for each sphere lit by each lighting direction (θ, ϕ) of the 32×32 mirror ball, we can compute a linear image $\hat{I}$ of each sphere under a novel lighting environment $\hat{L}$ as a linear combination of its basis images. Slicing the reflectance field into individual pixels $R_{x,y}(\theta, \phi)$, we generate $\hat{I}_{x,y}$ with (1), where $L_i(\theta, \phi)$ represents the color and intensity of light in the novel lighting environment for the direction (θ, ϕ):

$$\hat{I}_{x,y} = \sum_{\theta,\phi} R_{x,y}(\theta, \phi) L_i(\theta, \phi). \tag{1}$$

The network outputs Q, a log space image of omnidirectional HDR lighting in the mirror ball mapping, with pixel values $Q_i(\theta, \phi)$. Thus we render each sphere with (2):

$$\hat{I}_{x,y} = \sum_{\theta,\phi} R_{x,y}(\theta, \phi) e^{Q_i(\theta,\phi)}. \tag{2}$$

The ground truth sphere images I are LDR, 8-bit, gamma-encoded images, possibly with clipped pixels. Accordingly, we clip the rendered sphere images with a differentiable soft-clipping function Λ, $n = 40$:

$$\Lambda(p) = 1 - \tfrac{1}{n} \log\left(1 + e^{-n(p-1)}\right). \tag{3}$$

We then gamma-encode the clipped linear renderings with γ, to match I. We mask out the pixels in the corners of each ball image with a binary mask $\hat{M}$, producing the masked L_1 reconstruction loss L_{rec} for BRDFs $b = [0, 1, 2]$, where λ_b represents an optional weight for each BRDF:

$$L_{\text{rec}} = \sum_{b=0}^{2} \lambda_b \left\| \hat{M} \odot (\Lambda(\hat{I}_b)^{\frac{1}{\gamma}} - \Lambda(I_b)) \right\|_1. \qquad (4)$$

Adversarial loss: Minimizing only $\mathbf{E}[L_{\text{rec}}]$ produces blurred, low-frequency illumination. While this might be acceptable for lighting diffuse objects, rendering shiny objects with realistic specular reflections requires higher frequency lighting. Recent works in image inpainting and synthesis [36, 28, 58, 21, 57, 45] leverage Generative Adversarial Networks [16] for increased image detail, adding an adversarial loss to promote multi-modal outputs rather than a blurred mean of the distribution. We train our network in a similar framework to render *plausible* clipped mirror ball images, of which we have many real examples. This is perceptually motivated, as humans have difficulty reasoning about reflected light directions [37, 47], which digital artists leverage when environment mapping [5] reflective objects with arbitrary images. Furthermore, real-world lighting is highly regular, statistically [13].

Similar to Pathak et al. [36], we use an auxiliary discriminator network D with our base CNN as the generator G. During training, G tries to trick D, producing clipped mirror ball images appearing as "real" as possible. D tries to discriminate between real and generated images. We condition D on a few pixels from the original image surrounding the ball: we sample the four corners of the cropped ground truth mirror ball image, and bilinearly interpolate a 32×32 hallucinated background, as if the mirror ball were removed. We then softclip and composite both the ground truth and predicted mirror ball onto this "clean plate" with alpha blending (yielding I_c, $\hat{I}_c$) providing D with local color cues and ensuring that samples from both sets have the same perceptual discontinuity at the sphere boundary. Given input image x, G learns a mapping to Q, $G : x \to Q$, used to render a mirror ball with (2). The adversarial loss term, then, is:

$$\begin{aligned} L_{\text{adv}} &= \log D(\Lambda(I_c)) \\ &\quad + \log(1 - D(\Lambda(\textstyle\sum_{\theta,\phi} R(\theta,\phi) e^{G(x;\theta,\phi)})^{\frac{1}{\gamma}})). \qquad (5) \end{aligned}$$

Joint objective: The full objective is therefore:

$$G^* = \arg\min_G \max_D \, (1 - \lambda_{\text{rec}})\mathbf{E}[L_{\text{adv}}] + \lambda_{\text{rec}}\mathbf{E}[L_{\text{rec}}]. \qquad (6)$$

3.5. Implementation Details

We use TensorFlow [2] and train for 16 epochs using the ADAM [25] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a learning rate of 0.00015 for G, and, as is common, one 100× lower for D, alternating between training D and G. We set $\lambda_{\text{rec}} = 0.999$, with $\lambda_b = 0.2, 0.6, 0.2$ for the mirror, diffuse, and matte silver BRDFs respectively, and use $\gamma = 2.2$, as the camera's video mode employs image-dependent tone-mapping. We use a batch size of 32 and batch normalization [22] for all layers but the last of G and D. We use ReLU6 activations for G and ELU [9] for D. For our mobile demo (supplemental materials), we use TFLite. For data augmentation, we horizontally flip the input and ground truth images. We found that data augmentation by modifying white balance and exposure did not improve results, perhaps since they simulated unlikely camera responses.

Datasets: We collected 37.6 hours of training video using a Google Pixel XL mobile phone, in a variety of indoor and outdoor locations, times of day, and weather conditions, generating 4.06 million training examples. We bias the data towards imagery of surfaces or ground planes where one might want to place a virtual AR object. For test data, we collected 116 new one-minute videos (211.7k frames) with the same camera and separated them into four sets: unseen indoor and outdoor (UI, UO) and seen indoor and outdoor (SI, SO). "Unseen" test videos were recorded in new locations, while the "seen" were *new* videos recorded in previously-observed environments. We evaluate our method on the following videos: 28 UI (49.3k frames), 27 UO (49.7k frames), 27 SI (49.9k frames), and 34 SO (62.7k frames). Test data will be publicly released.

4. Evaluation

4.1. Quantitative Results

Accurate lighting estimates should correctly render objects with arbitrary materials, so we measure lighting accuracy first using L_{rec}, comparing with ground truth LDR spheres. We show the average per-pixel L_1 loss for each *unseen* test dataset for each material and the per-pixel linear RGB angular error θ_{rgb} for the diffuse ball, a distance metric commonly used to evaluate white-balance algorithms (see Hordley and Finlayson [19]), in Table 1 (top). (Minimizing θ_{rgb} during training did not improve results.) We show results for *seen* test sets in supplemental material.

Ablation studies: We assess the importance of the different loss terms, L_{rec} for each BRDF and L_{adv}, and report L_{rec} and θ_{rgb} for networks supervised using subsets of the loss terms in Table 1. Training with only the mirror BRDF or only the diffuse BRDF leads to higher L_{rec} for the others. However, training with only the matte silver BRDF still yields low L_{rec} for the diffuse sphere, suggesting they reveal similar lighting cues. In Fig. 5, we show the ground truth images and renderings produced for each loss variant. Visually, training with only the mirror ball $L_{1(m)}$ fails to recover the full dynamic range of lighting, as expected. Training with only the matte silver $L_{1(s)}$ or diffuse $L_{1(d)}$ fails to produce a realistic mirror ball; thus objects with sharp specular reflections could not be plausibly rendered. Training with L_{adv} yields higher frequency illumination as expected.

Table 1: Average L_1 loss by BRDF: diffuse (d), mirror (m), and matte silver (s), and RGB angular error θ_{rgb} for diffuse (columns), for our network trained with different loss terms (rows). We compare ground truth images with those *rendered* using our HDR lighting inference, for *unseen* indoor and outdoor locations.

	$L_{1(d)}$		$L_{1(s)}$		$L_{1(m)}$		$\theta^{\circ}_{\text{rgb}(d)}$	
Loss terms	UI	UO	UI	UO	UI	UO	UI	UO
$L_{1(m,d,s)} + L_{\text{adv}}$	0.12	0.13	0.13	0.13	0.17	0.16	9.8	10.8
$L_{1(m,d,s)}$	0.12	0.13	0.12	0.13	0.15	0.14	9.9	11.0
$L_{1(m)}$	0.20	0.18	0.16	0.15	0.14	0.13	11.0	13.5
$L_{1(s)}$	0.12	0.13	0.13	0.13	0.21	0.20	10.0	11.4
$L_{1(d)}$	0.12	0.13	0.15	0.15	0.28	0.27	10.0	11.2

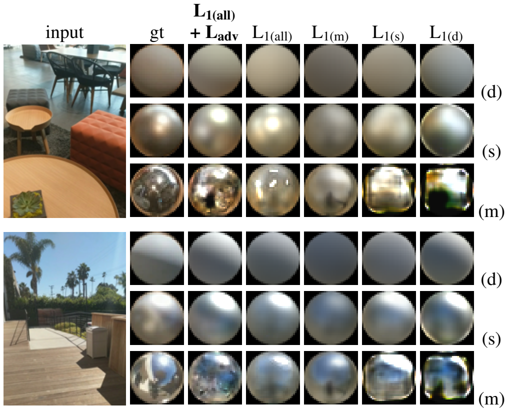


Figure 5: Ablation study: Unseen image inputs, ground truth, and rendered images of diffuse(d), matte silver(s) and mirror(m) spheres, lit with HDR lighting inference from networks trained using different loss terms(top). Our full method is labeled in bold.

4.2. Qualitative Results

Ground truth comparisons: In Fig. 6, we show examples of ground truth spheres compared with those rendered using image-based relighting and our HDR lighting inference, for each BRDF. These examples correspond to the 25th, 50th, and 75th percentiles for the L_{rec} loss.

Virtual object relighting: We 3D-print two identical bunnies using the model from [1]. The two are coated with paints of measured reflectance: diffuse gray (34.5% reflective) and matte silver (49.9% reflective), respectively. We photograph these "real" bunnies in different scenes using the Google Pixel XL, also capturing a clean plate for lighting inference and virtual object compositing. In Fig. 7 we compare the real bunny images (b, f) to off-line rendered composites using our lighting estimates (d, h) (IBL rendering described in the supplemental materials). We also record ground truth HDR lighting as in [10] using a Canon 5D Mark III, color correcting the raw linear HDR panorama so it matches the LDR phone image. We fit a linearization curve for each LDR input using a color chart, however the phone's image-dependent tone-mapping makes radiometric alignment challenging. We compare renderings using the ground truth and predicted lighting in Fig. 7 (c, g).

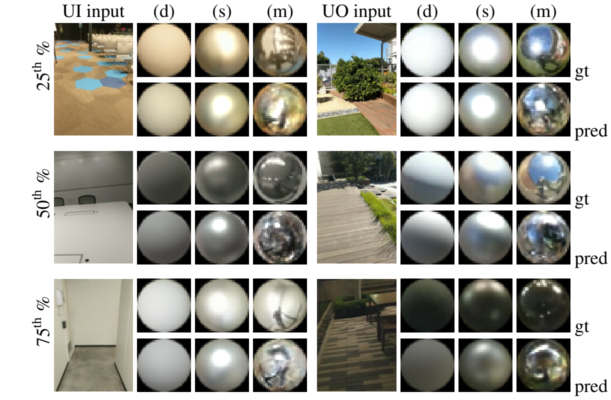


Figure 6: Qualitative comparisons between ground truth spheres and renderings using our HDR lighting inference and IBRL. Examples shown for 25th, 50th, and 75th percentiles for L_{rec}.

4.3. Comparisons with Previous Work

We retrain our network for the 3:4 aspect ratio input of the state-of-the-art methods for indoor [14] and outdoor [18] scenes, cropping a 1080×810 landscape image from the center of each portrait input and resizing to 192×144 to maintain our FC layer size. (Our comparison network thus observes half of the FOV of our standard network.) Gardner et al. [14] host a server to predict HDR lighting given an input image; Hold-Geoffroy et al. [18] also predict camera elevation. We randomly select 450 images from test sets UI and UO and retrieve their lighting estimates as HDR panoramas, converting them to the 32×32 mirror ball mapping and rotating them to camera space using the predicted camera elevation if given. We render spheres of each BRDF with IBRL and compare with ground truth, showing the average L_1 loss for each BRDF and θ_{rgb} for the diffuse ball in Table 2. We also show the relative error in total scene radiance measured by summing all diffuse sphere linear pixel values[1] in Table 9. We show comparison sphere renderings in Fig. 8 and bunny renderings in Fig. 7 (e, i), with more in supplemental materials along with a perceptual user study. We show significant improvements compared to both approaches, while requiring only one model that generalizes to both indoor and outdoor scenes. Without a specific sun and sky model, our network also infers diverse light sources for outdoor scenes. However, we present these results with two caveats: first, our training data are generated with a fixed FOV camera, which was varied and unknown for previous approaches, and second, our training and test

[1] Scene radiance is modulated by the albedo and foreshortening factor of the diffuse sphere, with greater frontal support, and we use $\gamma = 2.2$.

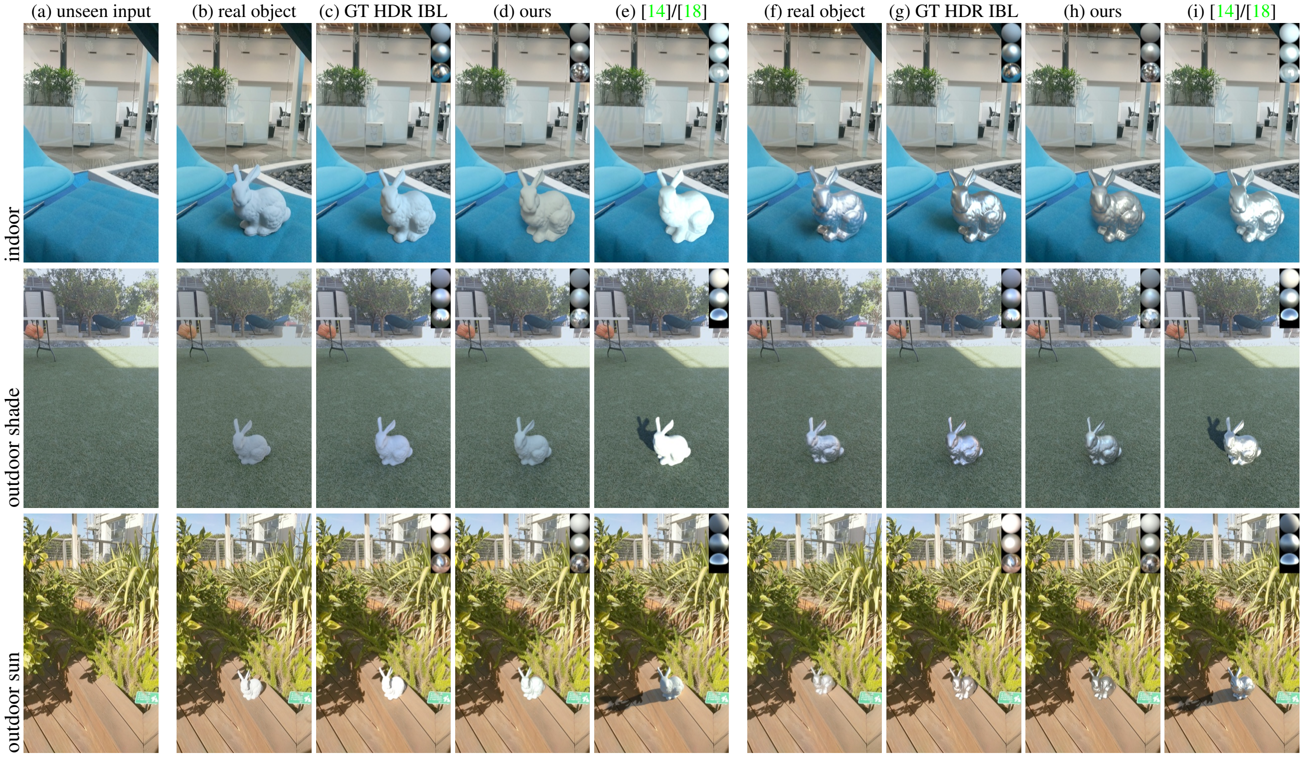


Figure 7: For each input image (a), we photograph a real 3D-printed bunny placed in the scene for two different BRDFs (b, f) and capture ground truth HDR panoramas at the bunny's location. Using GI rendering with IBL, we render a virtual bunny into the scene using ground truth lighting (c, g), our lighting inference (d, h), and that of the state-of-the-art methods for indoor [14] or outdoor [18] scenes (e, i).

Table 2: Quantitative comparisons with the previous state-of-the-art in indoor[14] and outdoor[18] lighting estimation. Average L_1 loss by BRDF: diffuse(d), mirror(m), and matte silver(s), and RGB angular error θ_{rgb} for the diffuse sphere. $n = 450$ for each.

	unseen indoor (UI)		unseen outdoor (UO)	
	ours	[14]	ours	[18]
$L_{1(d)}$	$\mathbf{0.13 \pm 0.07}$	0.21 ± 0.11	$\mathbf{0.13 \pm 0.08}$	0.25 ± 0.12
$L_{1(s)}$	$\mathbf{0.14 \pm 0.05}$	0.22 ± 0.06	$\mathbf{0.14 \pm 0.06}$	0.25 ± 0.07
$L_{1(m)}$	$\mathbf{0.18 \pm 0.03}$	0.23 ± 0.06	$\mathbf{0.17 \pm 0.04}$	0.34 ± 0.06
$\theta^{\circ}_{\text{rgb}(d)}$	$\mathbf{10.3 \pm 8.8^{\circ}}$	$11.9 \pm 7.2^{\circ}$	$\mathbf{11.2 \pm 10.9^{\circ}}$	$14.3 \pm 6.6^{\circ}$

data are generated with the same camera. Nonetheless, for mobile mixed-reality with a fixed-FOV, we show that optimizing for accurately rendered objects for multiple BRDFs improves lighting estimation.

Temporal consistency: We do not explicitly optimize for temporal consistency, but the adjacent video frames in our training data provide an indirect form of temporal regularization. In Fig. 10 we compare rendered results from four sequential frames for our approach and for that of Gardner et al. [14]. While we show qualitative improvement, adding a temporal loss term is of interest for future work.

4.4. Performance and Demonstration

Our inference runs at 12-20 fps on various mobile phone CPUs. We report performance for smaller networks and output lighting resolutions and timing for specific mobile phones in supplemental materials. We also authored a demo mobile application to predict lighting and render plausibly-lit virtual objects at interactive frame rates, using real-time

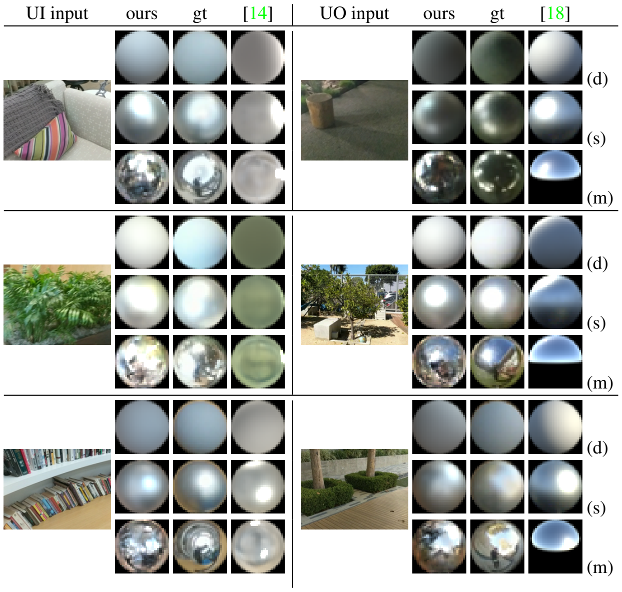


Figure 8: Ground truth and rendered spheres produced via IBRL using our predicted HDR lighting and that of the previous state-of-the-art for indoor [14] and outdoor [18] scenes.

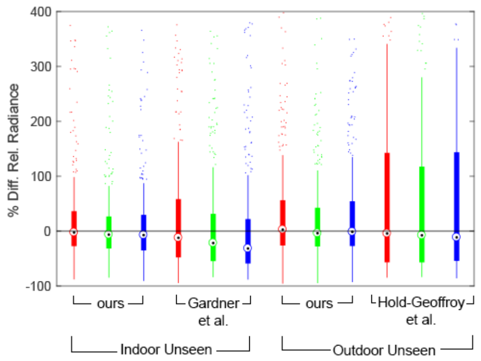
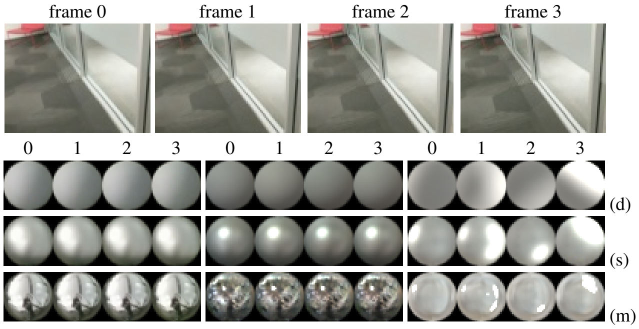


Figure 9: Boxplot of RGB relative radiance accuracy, measured by summing linear pixel values of the diffuse ball rendered with the HDR lighting estimates, and comparing with ground truth: (pred-gt)/gt, $n = 450$, for our approach and the previous state-of-the-art methods for indoor[14] and outdoor[18] scenes.



(a) ground truth (b) ours (c) [14]

Figure 10: Example ground truth spheres (a) and renderings produced with IBRL using our predicted illumination (b) and that of [14] (c), for four sequential UI video frames (top).

pre-computed radiance transfer [44] rendering.

5. Limitations and Future Work

Spatially-varying illumination: The reference spheres of the training data reflect the illumination from a point 60 cm in front of the camera and do not reveal spatially-varying lighting cues. Virtual AR objects are often placed on surfaces visible in the scene, and the light bouncing up from the surface should be the illumination on the object coming from below. A potential improvement to our technique would be to replace the bottom directions of our lighting estimate with pixel values sampled from the scene surface below each object, allowing objects placed in different parts of the scene to receive differently colored bounce light from their environments.

Using a different camera: Our test and training data are captured with the same camera. In Fig. 11 we show results for two images captured using a different mobile phone camera (Apple iPhone 6). Qualitatively, we observe differences in white balance, suggesting an avenue for future work. Similarly, our network is trained for a particular camera FOV and may not generalize to others.

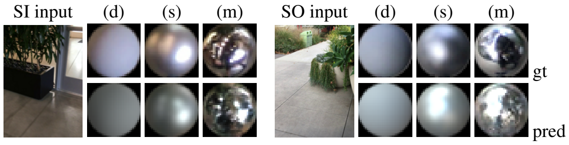


Figure 11: Example ground truth spheres and renderings produced with IBRL using our predicted HDR lighting, with input images from a different camera.

Challenging image content: Simple scenes lacking variation in surface normals and albedo (Fig. 12, left) can challenge our inference approach, and scenes dominated by a strongly hued material can also pose a challenge (Fig. 12, right). Adding knowledge of the camera's exposure and white balance used for each input image might improve the robustness of the inference.

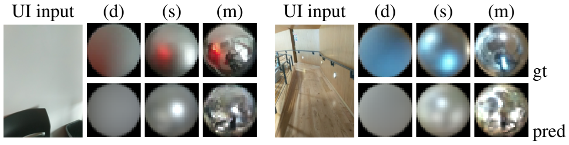


Figure 12: Example challenging scenes: ground truth spheres and renderings produced with IBRL using our predicted HDR lighting.

Future work: During mobile mixed reality sessions, objects are positioned on planes detected using sensor data fused with structure-from-motion [50]. Thus, computational resources are already devoted to geometric reasoning, which would be of interest to leverage for improved mixed reality lighting estimation. Furthermore, inertial measurements could be leveraged to continuously fuse and update lighting estimates as a user moves a phone throughout an environment. Similarly, as our training data already includes temporal structure, explicitly optimizing for temporal stability would be of interest. Lastly, one could increase generality by acquiring training data in a raw video format and simulating different camera models during training.

6. Conclusion

We have presented an HDR lighting inference method for mobile mixed reality, trained using only LDR imagery, leveraging reference spheres with different materials to reveal different lighting cues in a single exposure. This work is the first CNN-based approach that generalizes to both indoor and outdoor scenes for a single input image, with improved lighting estimation for mobile mixed reality as compared to previous work developed to handle only a single class of lighting.

References

[1] The Stanford 3D Scanning Repository. http://graphics.stanford.edu/data/3Dscanrep/. 6

[2] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 5

[3] J. T. Barron and J. Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2013. 2

[4] H. Barrow, J. Tenenbaum, A. Hanson, and E. Riseman. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2:3–26, 1978. 2

[5] J. F. Blinn and M. E. Newell. Texture and reflection in computer generated images. *Communications of the ACM*, 19(10):542–547, 1976. 5

[6] G. Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310:1–26, July 1980. 3

[7] 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. 3

[8] D. Cheng, J. Shi, Y. Chen, X. Deng, and X. Zhang. Learning scene illumination by pairwise photos from rear and front mobile cameras. In *Computer Graphics Forum*, volume 37, pages 213–221. Wiley Online Library, 2018. 3

[9] D. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations (ICLR)*, 2016. 5

[10] 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. 1, 2, 6

[11] P. Debevec, P. Graham, J. Busch, and M. Bolas. A single-shot light probe. In *ACM SIGGRAPH 2012 Talks*, page 10. ACM, 2012. 2

[12] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156. ACM Press/Addison-Wesley Publishing Co., 2000. 2, 3, 4

[13] R. O. Dror, A. S. Willsky, and E. H. Adelson. Statistical characterization of real-world illumination. *Journal of Vision*, 4(9):11–11, 2004. 5

[14] 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 Trans. Graph.*, 36(6):176:1–176:14, Nov. 2017. 2, 3, 6, 7, 8

[15] S. Georgoulis, K. Rematas, T. Ritschel, M. Fritz, L. Van Gool, and T. Tuytelaars. Delight-net: Decomposing reflectance maps into specular materials and natural illumination. *arXiv preprint arXiv:1603.08240*, 2016. 2

[16] 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. 5

[17] L. Gruber, T. Richter-Trummer, and D. Schmalstieg. Real-time photometric registration from arbitrary geometry. In *Mixed and Augmented Reality (ISMAR), 2012 IEEE International Symposium on*, pages 119–128. IEEE, 2012. 2

[18] Y. Hold-Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J.-F. Lalonde. Deep outdoor illumination estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, 2017. 2, 3, 6, 7, 8

[19] S. D. Hordley and G. D. Finlayson. Re-evaluating colour constancy algorithms. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 76–79. IEEE, 2004. 5

[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3

[21] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017. 5

[22] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015. 5

[23] 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. 3

[24] E. A. Khan, E. Reinhard, R. W. Fleming, and H. H. Bülthoff. Image-based material editing. In *ACM SIGGRAPH 2006 Papers*, SIGGRAPH '06, pages 654–663, New York, NY, USA, 2006. ACM. 2

[25] D. Kinga and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, volume 5, 2015. 5

[26] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan. Estimating natural illumination from a single outdoor image. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 183–190. IEEE, 2009. 3

[27] J.-F. Lalonde and I. Matthews. Lighting estimation in outdoor image collections. In *3D Vision (3DV), 2014 2nd International Conference on*, volume 1, pages 131–138. IEEE, 2014. 3

[28] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, page 4, 2017. 5

[29] G. Liu, D. Ceylan, E. Yumer, J. Yang, and J.-M. Lien. Material editing using a physically based rendering network. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2280–2288. IEEE, 2017. 2

[30] 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. 2

[31] J. Lopez-Moreno, E. Garces, S. Hadap, E. Reinhard, and D. Gutierrez. Multiple light source estimation in a single image. In *Computer Graphics Forum*, volume 32, pages 170–182. Wiley Online Library, 2013. 2

[32] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 183–194. Eurographics Association, 2007. 4

[33] D. Mandl, K. M. Yi, P. Mohr, P. M. Roth, P. Fua, V. Lepetit, D. Schmalstieg, and D. Kalkofen. Learning lightprobes for mixed reality illumination. In *Mixed and Augmented Reality (ISMAR), 2017 IEEE International Symposium on*, pages 82–89. IEEE, 2017. 2

[34] A. Meka, M. Maximov, M. Zollhoefer, A. Chatterjee, H.-P. Seidel, C. Richardt, and C. Theobalt. Lime: Live intrinsic material estimation. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[35] J. S. Nimeroff, E. Simoncelli, and J. Dorsey. Efficient re-rendering of naturally illuminated environments. In *Photorealistic Rendering Techniques*, pages 373–388. Springer, 1995. 2, 3

[36] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 5

[37] S. C. Pont and S. F. te Pas. Material illumination ambiguities and the perception of solid objects. *Perception*, 35(10):1331–1350, 2006. 5

[38] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 117–128. ACM, 2001. 2

[39] 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. 2, 4

[40] K. Rematas, T. Ritschel, M. Fritz, E. Gavves, and T. Tuytelaars. Deep reflectance maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4508–4516, 2016. 2

[41] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *arXiv preprint arXiv:1801.04381*, 2018. 3

[42] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. Sfsnet: learning shape, reflectance and illuminance of faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6296–6305, 2018. 3

[43] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5444–5453. IEEE, 2017. 3

[44] P.-P. Sloan, J. Kautz, and J. Snyder. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. *ACM Trans. Graph.*, 21(3):527–536, July 2002. 8

[45] 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. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*, pages 3847–3856. IEEE, 2018. 5

[46] J. Stumpfel, C. Tchou, A. Jones, T. Hawkins, A. Wenger, and P. Debevec. Direct hdr capture of the sun and sky. In *Proceedings of the 3rd international conference on Computer graphics, virtual reality, visualisation and interaction in Africa*, pages 145–149. ACM, 2004. 2, 4

[47] S. F. te Pas and S. C. Pont. A comparison of material and illumination discrimination performance for real rough, real smooth and computer generated smooth spheres. In *Proceedings of the 2nd Symposium on Applied Perception in Graphics and Visualization*, pages 75–81. ACM, 2005. 5

[48] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of Computer Vision and Pattern Recognition (CVPR 2018)*, 2018. 3

[49] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, page 5, 2017. 3

[50] S. Ullman. The interpretation of structure from motion. *Proc. R. Soc. Lond. B*, 203(1153):405–426, 1979. 8

[51] J. Unger, S. Gustavson, and A. Ynnerman. Densely sampled light probe sequences for spatially variant image based lighting. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia (GRAPHITE)*, GRAPHITE 06, pages 341–347. ACM, 2006. 2

[52] J. Waese and P. Debevec. P.: A real-time high dynamic range light probe. In *In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques: Conference Abstracts and Applications, p. 247. ACM Press/Addison-Wesley Publishing Co*, 2002. 2

[53] T. Y. Wang, T. Ritschel, and N. J. Mitra. Joint material and illumination estimation from photo sets in the wild. In *Proceedings of International Conference on 3DVision (3DV)*, 2018. selected for oral presentation. 2

[54] H. Weber, D. Prévost, and J.-F. Lalonde. Learning to estimate indoor lighting from 3d objects. *arXiv preprint arXiv:1806.03994*, 2018. 2

[55] 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. 2, 3

[56] Z. Xu, K. Sunkavalli, S. Hadap, and R. Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (TOG)*, 37(4):126, 2018. 3

[57] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017. 5

[58] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. *arXiv preprint*, 2018. 5

[59] Y. Yu, P. Debevec, J. Malik, and T. Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 215–224. ACM Press/Addison-Wesley Publishing Co., 1999. 2

[60] J. Zhang and J.-F. Lalonde. Learning high dynamic range from outdoor panoramas. In *IEEE International Conference on Computer Vision*, 2017. 3

[61] H. Zhou, J. Sun, Y. Yacoob, and D. W. Jacobs. Label denoising adversarial network (ldan) for inverse lighting of faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6238–6247, 2018. 3