A Dataset of Multi-Illumination Images in the Wild

A Dataset of Multi-Illumination Images in the Wild

Lukas Murmann[1] Michaël Gharbi[1, 2] Miika Aittala[1] Frédo Durand[1]

MIT CSAIL[1] Adobe Research[2]

Figure 1. Using our multi-illumination image dataset of over 1000 scenes, we can train neural networks to solve challenging vision tasks. For instance, one of our models can relight an input image to a novel light direction. Specular highlights pose a significant challenge for many relighting algorithms, but are handled gracefully by our network. Further analysis is presented in Section 4.2.

Abstract

Collections of images under a single, uncontrolled illumination [42] have enabled the rapid advancement of core computer vision tasks like classification, detection, and segmentation [26, 43, 18]. But even with modern learning techniques, many inverse problems involving lighting and material understanding remain too severely ill-posed to be solved with single-illumination datasets. The data simply does not contain the necessary supervisory signals. Multi-illumination datasets are notoriously hard to capture, so the data is typically collected at small scale, in controlled environments, either using multiple light sources [10, 53], or robotic gantries [8, 20]. This leads to image collections that are not representative of the variety and complexity of real-world scenes. We introduce a new multi-illumination dataset of more than 1000 real scenes, each captured in high dynamic range and high resolution, under 25 lighting conditions. We demonstrate the richness of this dataset by training state-of-the-art models for three challenging applications: single-image illumination estimation, image relighting, and mixed-illuminant white balance.

1. Introduction

The complex interplay of materials and light is central to the appearance of objects and to many areas of computer vision, such as inverse problems and relighting. We argue that research in this area is limited by the scarcity of datasets — the current data is often limited to individual samples captured in a lab setting, e.g. [8, 20], or to 2D photographs that do not encode the variation of appearance with respect to light [42]. While setups such as light stages, e.g. [10], can capture objects under varying illumination, they are hard to move and require the acquired object to be fully enclosed within the stage. This makes it difficult to capture everyday objects in their real environment.

In this paper, we introduce a new dataset of photographs of indoor surfaces under varying illumination. Our goal was to capture small scenes at scale (at least 1,000 scenes). We wanted to be able to bring the capture equipment to any house, apartment or office and record a scene in minutes. For this, we needed a compact setup. This appears to be at odds with the requirement that scenes be illuminated from different directions, since designs such as the light stage [10] have required a large number of individual lights placed around the scene. We resolved this dilemma by using indirect illumination and an electronic flash mounted on servos so that we can control its direction. As the flash gets rotated, it points to a wall or ceiling near the scene, which forms an indirect "bounce" light source. The reflected light becomes the primary illumination for the scene. We also place a chrome and a gray sphere in the scene as ground truth measurements of the incoming illumination.

Our capture process takes about five minutes per scene and is fully automatic. We have captured over a thousand scenes, each under 25 different illuminations for a total of 25,000 HDR images. Each picture comes segmented and labeled according to material. To the best of our knowledge, this is the first dataset of its kind: offering both everyday objects in context and lighting variations.

In Section 4, we demonstrate the generality and usefulness of our dataset with three learning-based applications: predicting the environment illumination, relighting single images, and correcting inconsistent white balance in pho-

tographs lit by multiple colored light sources.

We release the full dataset, along with our set of tools for processing and browsing the data, as well as training code and models.

2. Related Work

2.1. Multi-Illumination Image Sets

Outdoors, the sky and sun are natural sources of illumination varying over time. Timelapse datasets have been harvested both "in the wild" from web cameras [50, 46] or video collections [44], or using controlled camera setups [45, 29, 27].

Indoor scenes generally lack readily-available sources of illumination that exhibit significant variations. Some of the most common multi-illumination image sets are collections of flash/no-flash pairs [39, 12, 2]. These image pairs can be captured relatively easily in a brief two-image burst and enable useful applications like denoising, mixed-lighting white balance [22], or even BRDF capture [1]. Other applications, such as photometric stereo [51] or image relighting [10, 53], require more than two images for reliable results.

Datasets with more than two light directions are often acquired using complex hardware setups and multiple light sources [10, 20]. A notable exception, Mohan *et al*. [34] proposed a user-guided lighting design system that combines several illuminations of a single object. Like us, they acquire their images using a stationary motor-controlled light source and indirect bounce illumination, although within a more restrictive setup, and at a much smaller scale. For their work on user-assisted image compositing Boyadzhiev *et al*. [7] use a remote-controlled camera and manually shine a hand-held flash at the scene. This approach ties down the operator and makes acquisition times prohibitive (they report 20 minutes per scene). Further, handholding the light source makes multi-exposure HDR capture difficult. In contrast, our system, inspired by work of Murmann *et al*. [35], uses a motor-controlled bounce flash, which automates the sampling of lighting directions and makes multi-exposure HDR capture straightforward.

2.2. Material Databases

To faithfully acquire the reflectance of a real-world surface, one typically needs to observe the surface under multiple lighting conditions. The gold standard in appearance capture for materials is to exhaustively illuminate the material sample and photograph it under every pair of viewpoint and light direction, tabulating the result in a Bidirectional Texture Function (BTF). The reflectance values can then be read off this large table at render-time [8].

A variety of BTF datasets have been published [8, 30, 48], but the total number of samples falls far short of what is typically required by contemporary learning-based algorithms. A rich literature exists on simple, light-weight hardware capture systems [17], but the corresponding public datasets also typically contain less than a few dozen examples. Additionally, the scope, quality and format of these scattered and small datasets varies wildly, making it difficult to use them in a unified manner. Our portable capture device enables us to capture orders of magnitude more surfaces than existing databases and we record entire scenes at once —rather than single objects— "in the wild", outside the laboratory.

Bell *et al*. [5, 6] collected a large dataset of very loosely controlled photographs of materials from the Internet, enriched with crowd-sourced annotations on material class, estimated reflectance, planarity and other properties. Inspired by their approach, we collect semantic material class segmentations for our data, which we detail in section 3.3. Unlike ours, their dataset does not contain lighting variations.

Previous works have investigated the use of synthetic image datasets for material estimation [37, 49]. But even carefully crafted synthetic datasets typically do not transfer well to real scenes due remaining differences in scene complexity, object appearance, and image formation [40].

3. Dataset

Our dataset consists of 1016 interior scenes, each photographed under 25 predetermined lighting directions, sampled over the upper hemisphere relative to the camera. The scenes depict typical domestic and office environments. To maximize surface and material diversity, we fill the scenes with miscellaneous objects and clutter found in our capture locations. A selection of scenes is presented in Figure 2.

In the spirit of previous works [34, 35], our lighting variations are achieved by directing a concentrated flash beam towards the walls and ceiling of the room. The bright spot of light that bounces off the wall becomes a virtual light source that is the dominant source of illumination for the scene in front of the camera.

We can rapidly and automatically control the approximate position of the bounce light simply by rotating the flash head over a standardized set of directions (Figure 3). This alleviates the need to re-position a physical light source manually between each exposure [7, 32]. Our camera and flash system is more portable than dedicated light sources, which simplifies its deployment "in the wild".

The precise intensity, sharpness and direction of the illumination resulting from the bounced flash depends on the room geometry and its materials. We record these lighting conditions by inserting a pair of light probes, a reflective chrome sphere and a plastic gray sphere, at the bottom edge of every image [9]. In order to preserve the full dynamic range of the light probes and the viewed scene, all our pho-

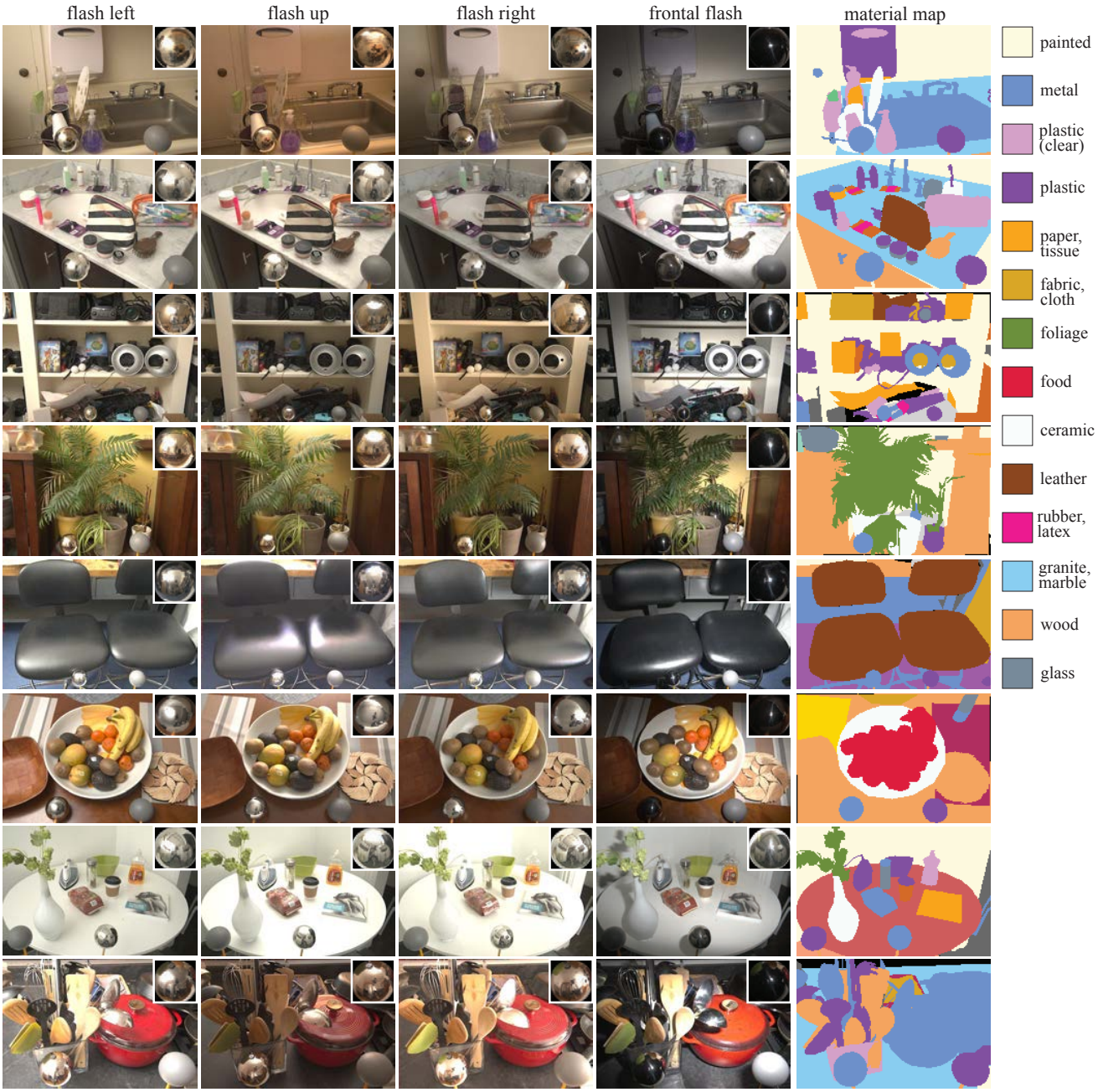

Figure 2. Eight representative scenes from our dataset. Each scene is captured under 25 unique light directions, 4 of which are shown in the figure. We strived to include a variety of room and material types in the dataset. Material types are annotated using dense segmentation masks which we show on the right.

tographs are taken with bracketed exposures.

As a post-process, we annotate the light probes, and collect dense material labels for every scene using crowdsourcing, as described in Section 3.3.

3.1. Image Capture

Our capture device consists of a mirrorless camera (*Sony α6500*), and an external flash unit (*Sony HVL-F60M*) which we equipped with two servo motors. The servos and camera are connected to a laptop, which automatically aims the flash and fires the exposures in a pre-programmed sequence. The 24mm lens provides a 52° horizontal and 36° vertical field of view.

At capture time, we rotate the flash in the 25 directions depicted in Figure 3, and capture a 3-image exposure stack for each flash direction. We switch off any room lights and

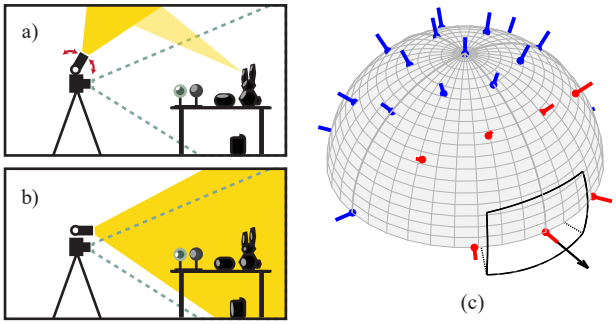

Figure 3. a) Most of our photographs are lit by pointing a flash unit towards the walls and the ceiling, creating a virtual bounce light source that illuminates the scene directionally. b) Some of the photographs are captured under direct flash illumination, where the beam of light intersects the field of view of the camera. c) The flash directions used in our dataset, relative to the camera viewing direction and frustum (black). The directions where direct flash illumination is seen in the view are shown in red, and the fully indirect ones in blue.

shut window blinds, which brings the average intensity of the ambient light to less than 1% of the intensity of the flash illumination. For completeness, we capture an extra, ambient-only, exposure stack with the flash turned off.

The 25 flash directions are evenly spaced over the upper hemisphere. In 18 of these directions, the cone of the flash beam falls outside the view of the camera, and consequently, the image is only lit by the secondary illumination from the bounce. In the remaining 7 directions, part or all of the image is lit by the flash directly. In particular, one of the directions corresponds to a typical frontal flash illumination condition.

Capturing a single set (78 exposures) takes about five minutes with our current setup. The capture speed is mostly constrained by the flash's recycling time (around 3.5 seconds at full power). Additional battery extender packs or high-voltage batteries can reduce this delay for short bursts. We found them less useful when capturing many image sets in a single session, where heat dissipation becomes the limiting factor.

3.2. HDR processing

The three exposures for each light direction are bracketed in 5-stops increments to avoid clipped highlights and excessive noise in the shadows. The darkest frame is exposed at $f/22$ ISO100, the middle exposure is $f/5.6$ ISO200, and the brightest image is recorded at $f/5.6$ ISO6400. The shutter speed is kept at the camera's fastest flash sync time, $1/160^{\text{th}}$ second to minimize ambient light. The camera sensor has 13 bits of useful dynamic range at ISO100 (9 bits at ISO6400). Overall, our capture strategy allows us to reconstruct HDR images with at least 20 bits of dynamic range.

Using the aperture setting to control exposure bracketing could lead to artifacts from varying defocus blur. We limit this effect by manually focusing the camera to the optimal depth, and by avoiding viewpoints with depth complexity beyond the depth-of-field that is achieved at $f/5.6$.

After merging exposures, we normalize the brightness of the HDR image by matching the intensity of the diffuse gray sphere. The gray sphere also serves as a reference point for white balance. This is especially useful in brightly-colored rooms that could otherwise cause color shifts.

3.3. Dataset Statistics

To ensure our data is representative of many real-world scenes, we collected images in 95 different rooms throughout 12 residential and office buildings, which allowed us to capture a variety of materials and room shapes.

In order to analyze the materials found throughout our dataset, we obtain dense material labels segmented by crowd workers, as shown in Figure 2 and 4. These annotations are inspired by the material annotations collected by Bell *et al*. [5], whose publicly available source code forms the basis of our annotation pipeline.

Figure 4 shows the distribution of materials in our data set. Specific room types have material distributions that differ markedly from the unconditioned distribution. For example, in kitchens we frequently find metal and wooden surfaces, but few fabrics (less than 5% of pixels). Bedrooms scenes on the other hand show fabrics in 38% of the pixels, but contain almost no metal surfaces (less than 4%).

4. Applications

In this section we present learning-based solutions to three long-standing vision problems: single-image lighting estimation, single-image relighting and mixed-illuminant white-balance. Our models are based on standard convolutional architectures, such as the U-net [41]. For all experiments, we normalize the exposure and white balance of our input images with respect to the gray sphere. We also mask out the chrome and gray spheres with black squares both at training and test time to prevent the networks from using this information directly.

4.1. Predicting Illumination from a Single Image

Single-frame illumination estimation is a challenging problem that arises e.g. when one wishes to composite a computer-generated object into a real-world image [9]. Given sufficient planning (as is common for visual effects in movies), illumination can be recorded at the same time the backdrop is photographed (e.g. by using a light probe). This is rarely the case for a posteriori applications. In particular, with the growing interest in augmented reality and mixed reality, the problem of estimating the illumination in uncontrolled scenes has received increased attention.

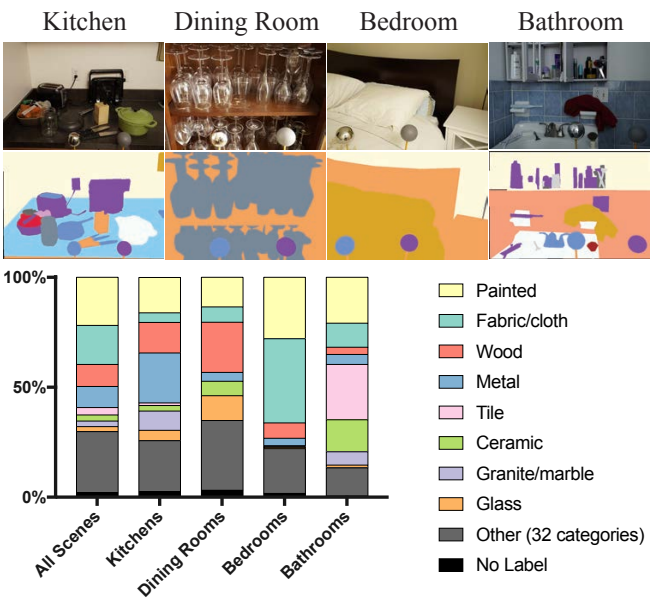

Figure 4. Crowd-sourced material annotations show that painted surfaces, fabrics, wood, and metal are the most frequently occurring materials in our dataset, covering more than 60% of all pixels. For some room types, the material distribution is markedly different from the average. For example, in kitchens we frequently encounter wood (shelves) and metal (appliances), bedroom scenes show a high frequency of fabrics, and the material distribution of bathrooms is skewed towards tiles and ceramics.

Several methods have explored this problem for outdoor images [28, 14, 15, 19, 31] as well as indoor environments [13]. Noting the lack of viable training data for indoor scenes, Gardner *et al*. explicitly detect light sources in LDR panoramas [52]. Our proposed dataset includes HDR light probes in every scene which makes it uniquely suitable for illumination prediction and other inverse rendering tasks [4] in indoor environments.

4.1.1 Model

We approach the single image illumination prediction problem by training a convolutional network on 256×256 image crops from our dataset. We ask the network to predict a 16×16 RGB chrome sphere, that we compare to our ground truth probe using an L_2 loss. The 256×256 input patch is processed by a sequence of convolution, ReLU, and Max-pooling layers, where we halve the spatial resolution and double the number of feature maps after each convolution. When the spatial resolution reaches 1×1 pixel, we apply a final, fully-connected layer to predict 768 numbers: these are reshaped into a 16×16 RGB light probe image. Exponentiating this images yields the final, predicted environment map. We provide the network details in supplemental material.

Figure 5. As the first application of our dataset, we train a deep network to predict environment maps from single input images. Our model consistently predicts the dominant light direction of the ground truth environment map. The model successfully estimates illumination based on shiny objects (a and g) and diffuse reflectors (e.g. row f). Rows h) and i) show failure cases where the network predicts low-confidence outputs close to the mean direction. We compare to Gardner *et al.* 's algorithm [13] which, while predicting visually plausible environment maps, lacks the precise localization of highlights shown by our technique. (Please ignore the vertical seam in Gardner *et al*.'s result. Their model uses a differing spherical parametrization, which we remap to our coordinate system for display.)

4.1.2 Compositing synthetic objects

Figure 5 shows some compositing results on a held-out test set. While our model does not always capture the finer color variations of the diffuse global illumination, its prediction of the dominant light source is accurate. Figure 7 shows one of the test scenes, with synthetic objects composited. The synthetic geometry is illuminated by our predicted environment maps and rendered with a path-tracer. Note that the ground truth light probes visible in the figure were masked during inference, and therefore *not* seen by our network.

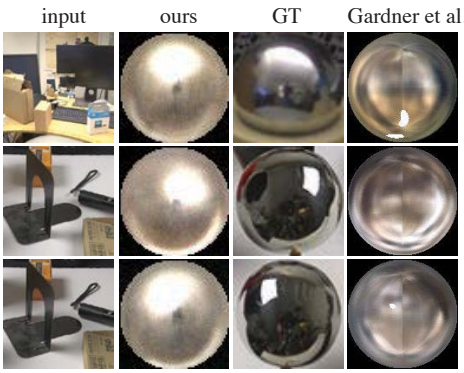

Figure 6. We validate our model's ability to generalize beyond bounce flash illumination. The top row show an office scene with regular room lights. The bottom two rows show a scene illuminated by softboxes, first lit from the right and then from the left. The second set of results suggests that our model can aggregate information from shadows to infer the light source position.

4.1.3 Evaluation

We evaluated our model on a held-out test subset of our data and compared it to a state-of-the-art illumination prediction algorithm by Gardner *et al.* [13]. Compared to their technique, our model more accurately predicts the direction of the bounce light source (see Figure 5). In comparison, Gardner *et al.*'s model favors smoother environment maps and is less likely to predict the directional component of the illumination. For visual comparison, we warp the 360° panoramas produced by their technique to the chrome sphere parameterization that is used throughout our paper.

In order to quantify the performance of the chrome sphere predictor, we analyzed the angular distance between the predicted and true center of the light source for 30 test scenes. Our technique achieves a mean angular error of 26.6° (std. dev. 10.8°), significantly outperforming Gardner *et al.*'s method, which achieves a mean error of 68.5° (std. dev. 38.4°). Visual inspection suggests that the remaining failure cases of our technique are due to left/right symmetry of the scene geometry, mirrors, or simply lack of context in the randomly chosen input crops (see Figure 5 bottom).

We verified that our model generalizes beyond bounce flash light sources using a small test set of pictures taken under general non-bounce flash illumination. The results of this experiment are presented in Figure 6.

4.2. Relighting

A robust and straightforward method to obtain a relightable model of a scene is to capture a large number of basis images under varying illumination, and render new images as linear combinations of the basis elements. Light stages [10] work according to this principle. With finitely many basis images, representing the high frequency content of a scene's light transport operator (specular highlights, sharp shadow boundaries, etc.) is difficult. Despite this fundamental challenge, prior work has successfully exploited the regularity of light transport in natural scenes to estimate the transport operator from sparse samples [36, 38, 47]. Recent approaches have employed convolutional neural networks for the task, effectively learning the regularities of light transport from synthetic training data and reducing the number of images required for relighting to just a handful [53].

Figure 7. We use the environment maps predicted by our model to illuminate virtual objects and composite them onto one of our test scenes. The light probe in the bottom of the frame shows the ground truth lighting. (Note that these probes are masked out before feeding the image to the network).

In our work, we demonstrate relighting results from a *single input image* on real-world scenes that exhibit challenging phenomena, such as specular highlights, self-shadowing, and interreflections.

4.2.1 Model

We cast single-image relighting as an image-to-image translation problem. We use a convolutional neural network based on the U-net [41] architecture to map from images illuminated from the left side of the camera, to images lit from the right (see supplemental material for details). Like in Section 4.1, we work in the log-domain to limit the dynamic range of the network's internal activations. We use an L_1 loss to compare the *spatial gradients* of our relit output to those of the reference image, lit from the right. We found this gradient-domain loss to yield sharper results. It also allows the network to focus more on fine details without being overly penalized for low-frequency shifts due to the global intensity scaling ambiguity (the left- and right-lit images might not have the same average brightness, depending on room geometry).

4.2.2 Evaluation

Our model faithfully synthesizes specular highlights, self- and cast-shadows, as well as plausible shading variations. Without a large-scale training dataset for an end-to-end solution, a valid straw man approach would be to use exist-

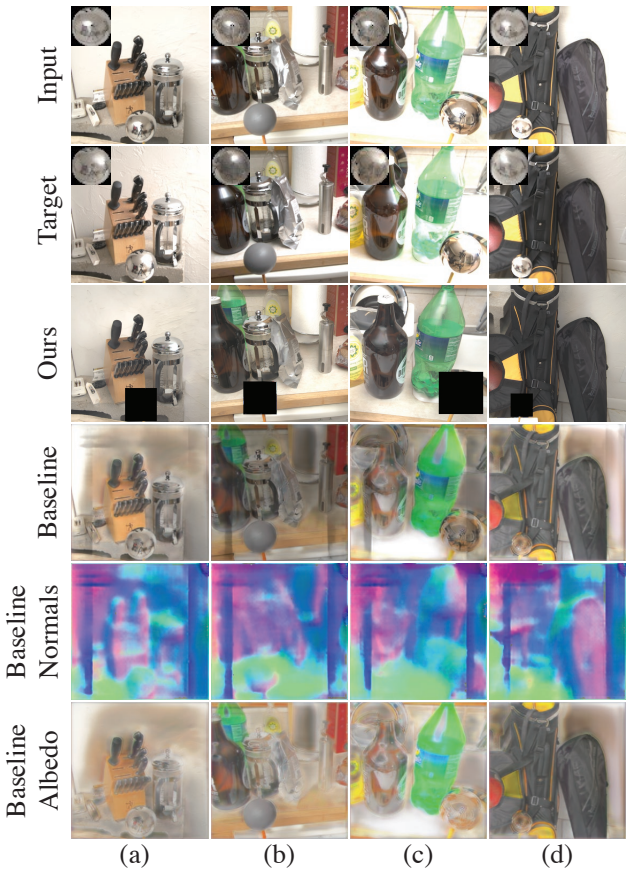

Figure 8. The second application of our data is learning image relighting using a single input image. The trained model synthesizes moving specular highlights (a, b, c) and diffuse shading (b, d), and correctly renders shadows behind occluders (d). For the baseline result, we first estimate normals and diffuse albedo using published models, and then re-render the image as lit by the target environment map.

ing techniques and decompose the input into components that can be manipulated for relighting (e.g. normals and albedo). We provide such a baseline for comparison. It uses a combination of deep single-image normal estimation [54] and learned intrinsic image decomposition [24] (see Figure 8). Both components are state-of-the-art in their respective fields and have source code and trained models publicly available.

This baseline illustrates the challenges in decomposing the single-image relighting problem into separate inverse rendering sub-problems. Specifically, major sources of artifacts include: incorrect or blurry normals, and incorrect surface albedo due to the overly simplistic Lambertian shading model. The normals and reflectance estimation networks were independently trained to solve two very difficult problems. This is a arguably more challenging than our end-to-end relighting application and, also unnecessary for plausible relighting.

Our end-to-end solution does not enforce this explicit geometry/material decomposition and yields far superior results. More relighting outputs produced by our model are shown in Figure 1 and in the supplemental material.

4.3. Mixed-Illumination White-Balance

White-balancing an image consists in neutralizing the color cast caused by non-standard illuminants, so that the photograph appears lit by a standardized (typically white) light source. White-balance is under-constrained, and is often solved by modeling and exploiting the statistical regularities in the colors of lights and objects. The most common automatic white balance algorithms make the simplifying assumption that the entire scene is lit by a single illuminant. See [16] for a survey. This assumption rarely holds in practice. For instance, an interior scene might exhibit a mix of bluish light (e.g. from sky illuminating the scene through a window) and warmer tones (e.g. from the room's artificial tungsten light bulbs). Prior work has formulated a local gray-world assumption to generalize white balance to the mixed-lighting case [11], exploiting the difference in light colors in shadowed vs. sunlit areas for outdoor scenes [25], or flash/no-flash image pairs [33, 21, 23].

Here again, we approach white-balancing as a supervised learning problem. Because our dataset contains high-dynamic range linear images with multiple lighting conditions, it can be used to simulate a wide range of new mixed-color illuminations by linear combinations. We exploit this property to generate a training dataset for a neural network that removes inconsistent color casts from mixed-illumination photographs.

4.3.1 Mixed-illuminant data generation

To create a training set of input/output pairs, we extract 256×256 patches from our scenes at multiple scales. For each patch, we choose a random number of light sources $n \in \{1, \ldots, 4\}$. Each light index corresponds to one of 25 available flash directions, selected uniformly at random without replacement. We denote by $I_1, \ldots, I_n$ the corresponding images.

For each light i, we sample its color with hue in $[0, 360]$, and saturation in $[0.5, 1]$, represented as a positive RGB gain vector α_i, normalized such that $||\alpha_i||_1 = 1$. We randomize the relative power of the light sources by sampling a scalar exposure gain g_i uniformly (in the log domain) between -3 and $+2$ stops. We finally assemble our mixed-colored input patch as the linear combination: $I = \frac{1}{n} \sum_{i=1}^{n} \alpha_i g_i I_i$.

We define the color-corrected target similarly, but without the color gains: $O = \frac{1}{n} \sum_{i=1}^{n} g_i I_i$.

Figure 9. The first row shows a mixed white-balance result on an image from a held-out test set. The input (left) is a linear combination of several (two here) illuminations of the same scene under varied color and intensity. The reference image (right) has the same energy but no color cast. Our output (middle) successfully removes the green and magenta shifts. The simple model we trained on our dataset, generalizes well to unseen, real RAW images (second row). The most noticeable failure case are skin tones (third row), which are entirely absent from our data set of static scenes.

4.3.2 Model

Like for the relighting problem, we use a simple convolutional network based on a U-net [41] to predict white-balanced images from mixed-lighting inputs (details in the supplemental).

To reduce the number of unknowns and alleviate the global scale ambiguity, we take the log transform of the input and target images, and decompose them in 2 chrominance u, v, and a luminance component l [3]:

$$u = \log(I^r + \epsilon) - \log(I^g + \epsilon), \quad (1)$$

$$v = \log(I^b + \epsilon) - \log(I^g + \epsilon), \quad (2)$$

$$l = \log(I^g + \epsilon), \quad (3)$$

where $\epsilon = 10^{-4}$, and the superscripts stand for the RGB color channels.

Our network takes as input u, v, l and outputs two correctly white-balanced chroma components. We assemble the final RGB output from l and the predicted chroma, using the reverse transform. Our model is trained to minimize an L_2 loss over the chroma difference.

4.3.3 Results

Our model successfully removes the mixed color cast on our test set and generalizes beyond, to real-world images. The main limitation of our technique is its poor generalization to skin tones, to which the human eye is particularly sensitive, but which are absent from our dataset of static indoor scenes. We present qualitative results in Figure 9 and in the supplemental video.

5. Limitations

A limitation of our capture methodology is that it requires good bounce surfaces placed not too far from the scene. This precludes most outdoor scenes and large indoor rooms like auditoriums. Our capture process requires the scene to remain static for several minutes, which keeps us from capturing human subjects. Compared to light stages or robotic gantries, the placement of our bounce light sources has more variability due to room geometry, and the bounce light is softer than hard lighting from point light sources. Finally, we only capture 25 different illuminations, which is sufficient for diffuse materials but under-samples highly specular ones.

6. Conclusions

We have introduced a new dataset of indoor object appearance under varying illumination. We have described a novel capture methodology based on an indirect bounce flash which enables, in a compact setup, the creation of virtual light sources. Our automated capture protocol allowed us to acquire over a thousand scenes, each under 25 different illuminations. We presented applications in environment map estimation, single-image relighting, and mixed white balance that can be trained from scratch using our dataset. We will release the code and data.

Acknowledgments

This work was supported in part by the DARPA REVEAL Program under Contract No. HR0011- 16-C-0030.

References

[1] Miika Aittala, Tim Weyrich, and Jaakko Lehtinen. Two-shot svbrdf capture for stationary materials. *ACM SIGGRAPH*, 2015. 2

[2] Yağız Aksoy, Changil Kim, Petr Kellnhofer, Sylvain Paris, Mohamed Elgharib, Marc Pollefeys, and Wojciech Matusik. A dataset of flash and ambient illumination pairs from the crowd. *ECCV*, 2018. 2

[3] Jonathan T. Barron. Convolutional color constancy. *ICCV*, 2015. 8

[4] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, 37(8):1670–1687, Aug 2015. 5

[5] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM SIGGRAPH*, 2013. 2, 4

[6] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. *CVPR*, 2015. 2

[7] Ivaylo Boyadzhiev, Sylvain Paris, and Kavita Bala. User-assisted image compositing for photographic lighting. *ACM SIGGRAPH*, 2013. 2

[8] Kristin J. Dana, Bram van Ginneken, Shree K. Nayar, and Jan J. Koenderink. Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics*, 1999. 1, 2

[9] Paul Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. *ACM SIGGRAPH*, 1998. 2, 4

[10] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. *ACM SIGGRAPH*, 2000. 1, 2, 6

[11] Marc Ebner. Combining white-patch retinex and the gray world assumption to achieve color constancy for multiple illuminants. In *Joint Pattern Recognition Symposium*, pages 60–67. Springer, 2003. 7

[12] Elmar Eisemann and Frédo Durand. Flash photography enhancement via intrinsic relighting. *ACM SIGGRAPH*, 2004. 2

[13] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM SIGGRAPH Asia*, 2017. 4, 5, 6

[14] Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Tinne Tuytelaars, and Luc Van Gool. Natural illumination from multiple materials using deep learning. *CoRR*, 2016. 4

[15] Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Efstratios Gavves, Mario Fritz, Luc Van Gool, and Tinne Tuytelaars. Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 4

[16] Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. Computational color constancy: Survey and experiments. *IEEE Transactions on Image Processing*, 2011. 7

[17] Dar'ya Guarnera, Giuseppe Claudio Guarnera, Abhijeet Ghosh, Cornelia Denk, and Mashhuda Glencross. BRDF Representation and Acquisition. *Computer Graphics Forum*, 2016. 2

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *ICCV*, 2017. 1

[19] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. *CVPR*, 2017. 4

[20] Michael Holroyd, Jason Lawrence, and Todd Zickler. A coaxial optical scanner for synchronous acquisition of 3D geometry and surface reflectance. *ACM SIGGRAPH*, 2010. 1, 2

[21] Zhuo Hui, Aswin C Sankaranarayanan, Kalyan Sunkavalli, and Sunil Hadap. White balance under mixed illumination using flash photography. *IEEE International Conference on Computational Photography*, 2016. 7

[22] Zhuo Hui, Kalyan Sunkavalli, Sunil Hadap, and Aswin C. Sankaranarayanan. Post-capture lighting manipulation using flash photography. *CoRR*, 2017. 2

[23] Zhuo Hui, Kalyan Sunkavalli, Sunil Hadap, and Aswin C Sankaranarayanan. Illuminant spectra-based source separation using flash photography. *CVPR*, 2018. 7

[24] Carlo Innamorati, Tobias Ritschel, Tim Weyrich, and Niloy J. Mitra. Decomposing single images for layered photo retouching. *Computer Graphics Forum*, 2017. 7

[25] Rei Kawakami, Katsushi Ikeuchi, and Robby T Tan. Consistent surface color for texturing large objects in outdoor scenes. *ICCV*, 2005. 7

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012. 1

[27] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM SIGGRAPH*, 2014. 2

[28] Jean-François Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. Estimating the natural illumination conditions from a single outdoor image. *International Journal of Computer Vision*, 2011. 4

[29] Jean-François Lalonde and Iain Matthews. Lighting estimation in outdoor image collections. *International Conference on 3D Vision*, 2014. 2

[30] Jason Lawrence, Aner Ben-Artzi, Christopher DeCoro, Wojciech Matusik, Hanspeter Pfister, Ravi Ramamoorthi, and Szymon Rusinkiewicz. Inverse shade trees for non-parametric material representation and editing. *ACM SIGGRAPH*, 2006. 2

[31] Stephen Lombardi and Ko Nishino. Reflectance and illumination recovery in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. 4

[32] Vincent Masselus, Philip Dutré, and Frederik Anrys. The free-form light stage. Technical report, Departement of Computer Science, KU Leuven, 2002. 2

[33] Ryo Matsuoka, Tatsuya Baba, Mia Rizkinia, and Masahiro Okuda. White balancing by using multiple images via intrinsic image decomposition. *IEICE Transactions on Information and Systems*, 2015. 7

[34] Ankit Mohan, Reynold Bailey, Jonathan Waite, Jack Tumblin, Cindy Grimm, and Bobby Bodenheimer. Tabletop computed lighting for practical digital photography. *IEEE Transactions on Visualization and Computer Graphics*, 2007. 2

[35] Lukas Murmann, Abe Davis, Jan Kautz, and Frédo Durand. Computational bounce flash for indoor portraits. *ACM SIGGRAPH Asia*, 2016. 2

[36] Shree K. Nayar, Peter N. Belhumeur, and Terry E. Boult. Lighting sensitive display. *ACM Transactions on Graphics*, 2004. 6

[37] Keunhong Park, Konstantinos Rematas, Ali Farhadi, and Steven M Seitz. Photoshape: photorealistic materials for large-scale shape collections. *ACM SIGGRAPH Asia*, 2018. 2

[38] Pieter Peers, Dhruv K. Mahajan, Bruce Lamond, Abhijeet Ghosh, Wojciech Matusik, Ravi Ramamoorthi, and Paul Debevec. Compressive light transport sensing. *ACM Transactions on Graphics*, 2009. 6

[39] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. Digital photography with flash and no-flash image pairs. *ACM SIGGRAPH*, 2004. 2

[40] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. *ECCV*, 2016. 2

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, 2015. 4, 6, 8

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015. 1

[43] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, 2013. 1

[44] Yichang Shih, Sylvain Paris, Frédo Durand, and William T Freeman. Data-driven hallucination of different times of day from a single outdoor photo. *ACM SIGGRAPH Asia 2013*, 2013. 2

[45] Jessi Stumpfel, Chris Tchou, Andrew Jones, Tim Hawkins, Andreas Wenger, and Paul Debevec. Direct hdr capture of the sun and sky. *International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*, 2004. 2

[46] Kalyan Sunkavalli, Fabiano Romeiro, Wojciech Matusik, Todd Zickler, and Hanspeter Pfister. What do color changes reveal about an outdoor scene? *CVPR*, 2008. 2

[47] Jiaping Wang, Yue Dong, Xin Tong, Zhouchen Lin, and Baining Guo. Kernel nyström method for light transport. *ACM SIGGRAPH*, 2009. 6

[48] Michael Weinmann, Juergen Gall, and Reinhard Klein. Material classification based on training data synthesized using a btf database. In *European Conference on Computer Vision (ECCV)*, pages 156–171, 2014. 2

[49] Michael Weinmann, Juergen Gall, and Reinhard Klein. Material classification based on training data synthesized using a btf database. *ECCV*, 2014. 2

[50] Yair Weiss. Deriving intrinsic images from image sequences. *ICCV*, 2001. 2

[51] Robert J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 1980. 2

[52] Jianxiong Xiao, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. *CVPR*, 2012. 5

[53] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics*, 2018. 1, 2, 6

[54] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. *CVPR*, 2017. 7