Physically-inspired Deep Light Estimation from a Homogeneous-Material Object for Mixed Reality Lighting

Jinwoo Park, Hunmin Park, *Student Members, IEEE*, Sung-eui Yoon, and Woontack Woo, *Members, IEEE*

Fig. 1: Given a single LDR image of a reference object with a homogeneous specular or diffuse material, the proposed method estimates HDR illumination of a scene. Without additional material or shape information, it predicts accurate directions and colors of incident radiances in the form of an environment map. In each example scene, the reconstructed HDR illumination (tone-mapped for visualization) can be used to relight a virtual object coherently with the scene for realistic mixed reality. In the indoor (left) and outdoor (right) scenes, illumination is estimated from the elephant statue and the spherical statue, respectively.

Abstract— In mixed reality (MR), augmenting virtual objects consistently with real-world illumination is one of the key factors that provide a realistic and immersive user experience. For this purpose, we propose a novel deep learning-based method to estimate high dynamic range (HDR) illumination from a single RGB image of a reference object. To obtain illumination of a current scene, previous approaches inserted a special camera in that scene, which may interfere with user's immersion, or they analyzed reflected radiances from a passive light probe with a specific type of materials or a known shape. The proposed method does not require any additional gadgets or strong prior cues, and aims to predict illumination from a single image of an observed object with a wide range of homogeneous materials and shapes. To effectively solve this ill-posed inverse rendering problem, three sequential deep neural networks are employed based on a physically-inspired design. These networks perform end-to-end regression to gradually decrease dependency on the material and shape. To cover various conditions, the proposed networks are trained on a large synthetic dataset generated by physically-based rendering. Finally, the reconstructed HDR illumination enables realistic image-based lighting of virtual objects in MR. Experimental results demonstrate the effectiveness of this approach compared against state-of-the-art methods. The paper also suggests some interesting MR applications in indoor and outdoor scenes.

Index Terms—Light estimation, light probe, physically-based rendering, deep learning, coherent rendering, mixed reality.

1 Introduction

In mixed reality (MR), photometric registration enables coherent lighting of virtual objects in a real-world environment, which consequently provides a more immersive experience for users. For this reason, it is necessary to obtain the illumination of a scene, which usually involves a specialized gadget for capturing the whole environment. For example, a fish-eye lens or omnidirectional cameras are generally inserted in a scene of interest [27, 30], and they might interrupt users' immersion. Using a spherical light probe such as a chrome ball is one of the pioneering works aimed at directly inferring incident illumination by observing the reflected lights from it [10]. However, a chrome ball is a special setup, which does not commonly appear in casual scenes.

To substitute those approaches, there are analytical methods to infer illumination by solving a challenging inverse rendering problem from single or multiple images with a limited field of view (FoV). Color and, more specifically, radiance reflected on a surface point of the object go through physical interactions among incident lights, the object's material properties, and geometrical structure, which are described in the rendering equation [45]. Thus, estimating the light information from the remaining unknown properties is a complex ill-posed problem. For this reason, previous studies employed various types of prior conditions such as known shapes or constrained materials including Lambertian surfaces for diffuse shading. However, using a predefined shape restricts a casual MR experience [1], and obtaining arbitrary shape information by reconstruction might require expensive preprocessing [40]. Furthermore, it is challenging to infer the high-frequency illumination only from diffuse surfaces, although this assumption has a great advantage in terms of simplifying the rendering equation.

In this work, a recent learning-based approach is employed to effectively overcome those strict assumptions. With a large dataset, deep learning has successfully shown its capacity to estimate or extract meaningful features from a single image, such as semantic classification [32], object segmentation [8], and a reflectance map [39]. In addition, thanks to its multi-layered structure helpful to handle a complex non-linear problem with unconstrained input images, deep learning has shown convincing solutions for ill-posed light estimation under weak constraints [15, 21, 38]. For this main reason, the proposed method follows the well-used composition of convolutional and deconvolutional layers but estimates HDR illumination from a single LDR input image of a reference object with a wide range of materials, shapes and environments. It is only assumed that the reference object has a uniform material and its shape does not have a dominantly concave structure that generates high occlusions. Furthermore, deep learning makes our method available in real-time as shown in our application (Sect. 5). Finally, the advantage of the proposed learning-based method can be

- *Jinwoo Park and Woontack Woo are with KAIST UVR Lab.. E-mail: {jinwooaa | wwoo}@kaist.ac.kr*
- *Hunmin Park and Sung-eui Yoon are with KAIST. E-mail: {95phm | sungeui}@kaist.ac.kr*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

maximized in MR applications where the only resource that can be used is an RGB image of a real-world scene with unknown objects.

Departing from prior approaches, our key idea is that the complex rendering equation is transformed into an approximate, yet simpler and more effective form that factorizes irradiance information from other factors. The main benefit of this factorization is that it is possible to design our first network that learns to extract only the irradiance term on each surface point of the reference object. This network helps in simplifying the challenging estimation problem by decreasing dependency on material properties, including diffuse albedo and specular reflectivity, in the early stage. As a result, instead of directly inferring the illumination from a raw input object, this intermediate task provides more robust results for unknown specular or diffuse materials.

Based on the previously refined input object, LDR and HDR illuminations are estimated through the second and third networks, respectively. The second network maps the estimated irradiance on the object surfaces onto an LDR environment map that serves as a distant light map. Although this is a highly challenging domain transformation problem without an object shape, it can be handled more reliably over prior approaches by employing the gradual training method proposed in this study. From the inferred LDR environment map, the third network estimates HDR radiances in saturated areas such as light sources. Finally, those estimated LDR and HDR values are jointly used in the linear interpolation function to reconstruct the final HDR radiance map.

Through our step-by-step approach to the complex regression problem, the reconstructed HDR illumination can be practically utilized for coherent rendering in MR applications. Furthermore, the proposed method can be robust to various conditions in real-world applications thanks to our synthetic training dataset where millions of objects are photo-realistically rendered in different environments without using a real dataset that involves laborious obtaining processes.

Overall, the main contributions and benefits of the present work are summarized as follows:

- A novel learning-based method with sequential end-to-end deep neural networks that are physically inspired for estimating omnidirectional HDR illumination from a single LDR image of a reference object with a limited FoV.
- An effective approach that gradually decreases dependency on the unknown properties of an observed object to cover a wide range of uniform materials and shapes, which makes the proposed method more useful in casual MR applications.
- A new synthetic dataset where millions of objects are rendered in a physically correct manner with diverse materials, shapes, and environments, including indoor and outdoor conditions.

2 Related Work

In this section, we discuss direct and indirect light estimation methods, as well as recent deep learning approaches.

Direct Approaches Using Image Sensors. One of the traditional methods for directly obtaining illumination as a distant light map is capturing it by means of a camera. For example, an environment map can be obtained by placing a camera with a wide-angle lens like a fish-eye lens inside a scene [27, 30]. Although this is the simplest and most exact method, illumination can be obtained only at a fixed position where the camera is placed. Furthermore, inserting an additional gadget in the scene may restrict the spatial scope of user activity in MR.

To solve this issue, other direct methods only use a casual camera attached to a mobile device for MR and estimate or approximate the lighting condition based on image-processing. Kan [25] proposed a practical way to catch environmental illumination in HDR by composing multiple images taken from a casual mobile device with limited FoV while taking much time to synthesize one environment map. Schwandt and Broll [42] proposed the creation of a plausible environment map from a single RGB image by splitting and approximately mapping it onto equirectangular representation.

Differently from those direct methods, in this study, no additional devices are inserted into the scene, and no post-processing is used. Instead, HDR illumination is inferred from a single image of a reference object through learned regression.

Indirect Methods from Light Probes. Without additional cameras inserted into a real scene, various light estimation methods observing reference objects have been proposed. As a pioneering work, Debevec's method [10] has been widely used to capture HDR illumination by analyzing reflected light from a mirror ball-like light probe. Although Pessoa et al. [37] rendered physically plausible results by using this method, getting an HDR radiance map required an offline process that cannot be used in interactive applications. Aittala [1] presented a more interactive system using a diffuse ball as a light probe, but spherical harmonics (SH) cannot estimate high-frequency illumination, and the light probe was limited to the spherical shape.

In a different direction, more complex shapes of light probes were adopted [3, 33] with their geometrical information to analytically solve the ill-posed problems. Gruber et al. [19] used an arbitrary scene as a light probe and estimated the distant light based on SH coefficients. Similarly, Richter-Trummer et al. [40] used an arbitrarily shaped object reconstructed by a casual RGB-D sensor as a light probe and showed plausible relighting results by using estimated illumination, also in the SH form. Unfortunately, the light probe had to be reconstructed in a preprocessing step, and it required high computational power. Interestingly, Knorr and Kurz [31] showed that a human face could be a good light probe in a dynamic lighting condition. It also employed a low-order radiance transfer function [43] to get diffuse lights.

In contrast to the above methods, this study focuses on estimating the higher-frequency illumination than low-order SH representations from a light probe, while covering a wide range of materials and shapes, without any precomputation for reconstruction. Note that we only use a single LDR image to infer the HDR illumination through deep neural networks without requiring additional images in different exposures.

Deep Learning-based Light Estimation. Based on the convolutional neural network (CNN) [32] and deconvolution architecture [44, 47], recent works have suggested estimation methods that decompose a typical RGB image into its intrinsic components, e.g., natural illumination, material properties, or surface normal vectors. As one of the early works using CNN, Mandl et al. [35] learned a known object as a light probe in uniformly distributed camera views to infer 16 SH coefficients. Rematas et al. [39] estimated a reflectance map [22] from an LDR image of a specular light probe in a known class. In their following work [38], the estimated reflectance map provided Phong reflectance parameters and HDR illumination. Georgoulis et al. [16] inferred omnidirectional HDR illumination from partial reflectance maps of multiple materials and a background RGB image under the assumption that the material was fixed to the specular one.

Moreover, Meka et al. [36] proposed a physically motivated network that gradually decomposes the Blinn-Phong lighting model [4] to get diffuse and specular albedo values with a shininess parameter. This approach also tried to catch illumination based on well estimated material properties, but that was only possible when the shape information was provided. In addition, the assumption of white specular albedo caused the estimation of gray illumination. Other works inferred the indoor illumination [15] and sky parameters for simulating an outdoor environment map [21]. Their strength consisted in using only an LDR input image of an arbitrary scene even without a particular light probe while they were fitted to specific environments. Calian et al. [6] used a single LDR image of a human face to estimate extremely high-frequency outdoor illumination although they assumed the input had spatially constant albedo. Kan et al. [26] estimated a dominant light direction with temporal consistency from a general scene, but their method required an RGB-D image and had difficulty in inferring incident HDR lights.

In our work, a reference object is observed in diverse environments for the purpose of estimating indoor and outdoor illuminations. In addition, thanks to the proposed sequential steps of decreasing material and shape dependency, trained on our large photo-realistic dataset, illumination can be robustly estimated by covering various types of homogeneous materials and shapes without additional inputs such as a depth map or surface normals.

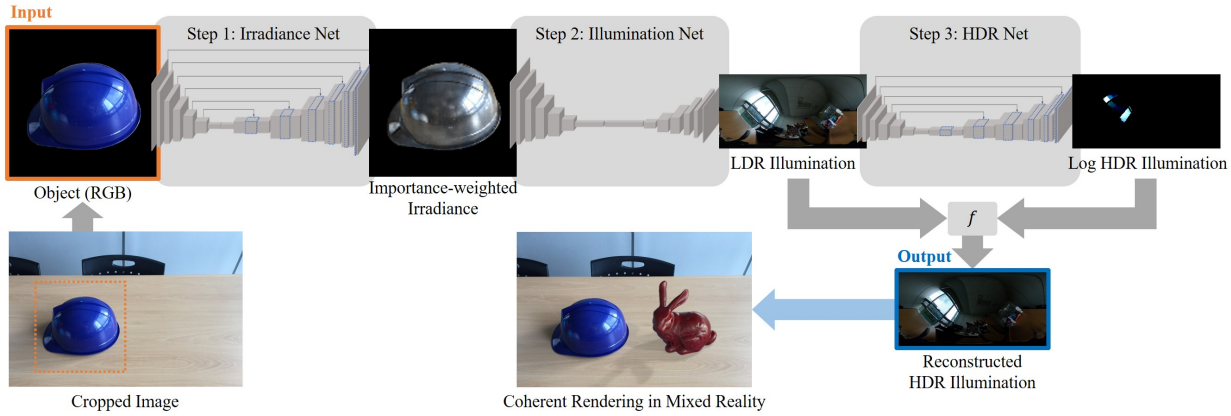


Fig. 2: Overview. Starting from a segmented, input RGB image of a reference object captured by a casual camera used in MR, we estimate an irradiance term, named as *Importance-weighted Irradiance*, on each object surface point, followed by estimating LDR and log HDR illumination. Finally, HDR radiance values are reconstructed by combining estimated LDR and log HDR illumination, based on the linear interpolation function f. This result enables MR applications to render photo-realistic virtual objects coherently with the real-world environment.

3 Physically-Inspired Light Estimation

In this paper, we propose a novel deep learning-based method that estimates HDR illumination from a single RGB image of a reference object with a homogeneous material. In the first trial, we found that learning illumination directly from a raw input object with unknown material and shape properties was still problematic. As a more effective way to solve this ill-posed inverse rendering problem, our main idea is to use a step-by-step approach, which divides such direct estimation into three smaller processes: extracting only the irradiance term on each surface point (Sect. 3.2) and predicting LDR illumination from that term (Sect. 3.3), followed by reconstructing HDR illumination (Sect. 3.4). All of those sequential networks are robustly trained on our physically-based synthetic dataset (Sect. 3.5) that covers a huge variety of materials, shapes of reference objects and lighting conditions.

In the overall process, the proposed method takes a cropped image with a segmented object as an input and estimates an HDR environment map as a final output, as shown in Fig. 2. Note that segmentation of a reference object can be done using various approaches [8,36]. Clipping a depth range of an input image stream from a depth sensor is one of the real-time methods to segment an object of interest for MR [7,13].

Above all, as our networks are motivated by the physical process of light transport from light sources into RGB values through reflections and refractions, the rendering equation [24] and our assumptions used in the equation are introduced. Then, we briefly explain the Cook-Torrance bidirectional reflectance distribution function (BRDF) [9], one of the most commonly used BRDFs for realistic rendering (Sect. 3.1).

3.1 Background of Physically-based Rendering

Colors or radiance values of an image can be explained by the rendering equation that consists of various factors such as a BRDF, the geometry, and incident radiances at measurement points. Thus, light estimation from a reference object requires solving the ill-posed problem.

To tackle this inverse rendering problem, we propose a deep learning-based regression method that employs a physically-inspired rendering model with Cook-Torrance BRDF. Although the Blinn-Phong model has been adopted in prior methods [36,38] due to its simple composition, we use a more realistic model. One of the main benefits of using it is that millions of photo-realistic images can be generated for training, instead of relying on a tedious and manual process of taking real images.

To understand the light estimation steps of our method, we briefly introduce the rendering equation [24] first:

$$L_o(p,\omega_o) = L_e(p,\omega_o) + \int_{\Omega^+} f_r(p,\omega_i,\omega_o) L_i(p,\omega_i)(\omega_i \cdot n) d\omega_i, \quad (1)$$

where L_o is the outgoing radiance that finally reaches the viewer's eye in an outgoing direction ω_o from surface point p. In addition, L_e is emitted radiance, if there is one. The integral part represents the sum of reflected radiances from surface point p in direction ω_o. In this part, L_i is the incident radiance from direction ω_i, scaled by a positive geometrical factor considering the angle between ω_i and normal vector n on the surface point. Again, L_i is scaled by BRDF f_r, which renders the reflectance ratio of incident light depending on the material. Finally, all of the scaled lights are accumulated over the upper hemisphere Ω^+.

In our method, it is assumed that the observed objects are not light-emitting, so L_e can be neglected in the equation. In addition, to simulate incoming lights, an image-based lighting method [10] with a distant light map is adopted, not taking account of any occlusions. Therefore, regardless of surface point p, they have different radiance values according to their incoming directions. Finally, it is assumed that an observed object has a homogeneous material without translucency.

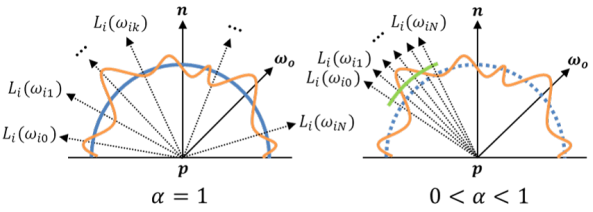
Regarding the reflectance model, we propose to use Cook-Torrance BRDF to support various realistic materials for generating high-quality synthetic images. Commonly, BRDF can be divided into the diffuse and specular parts [9]:

$$f_r(p,\omega_i,\omega_o) = k_d \frac{\rho}{\pi} + \underbrace{\frac{DFG}{4(\omega_o \cdot n)(\omega_i \cdot n)}}_{\text{Cook-Torrance BRDF}}, \quad (2)$$

where k_d is the coefficient for diffuse reflection. It can be obtained by $1-F$, where F is the Fresnel function that gives the ratio of the reflected light for specular reflection. Since Lambertian reflection is applied to diffuse shading, f_r for diffuse reflection is represented by $\frac{\rho}{\pi}$, where ρ denotes the albedo of diffuse reflection. Based on a microfacet model, Cook-Torrance BRDF for specular reflection is composed of three different functions: distribution function D, geometry function G, and Fresnel function F. Although the distribution and geometry functions are related to the geometric aspects of microfacets, such as surface roughness, the Fresnel function describes their optical property. More details of those functions are kindly explained in [2].

Finally, when the BRDF model is plugged into the original rendering equation, the following equation is obtained:

$$L_o(p,\omega_o) = \underbrace{k_d \frac{\rho}{\pi} \int_{\Omega^+} L_i(\omega_i)(\omega_i \cdot n) d\omega_i}_{\text{diffuse shading}} + \underbrace{\int_{\Omega^+} \frac{DFG}{4(\omega_o \cdot n)(\omega_i \cdot n)} L_i(\omega_i)(\omega_i \cdot n) d\omega_i}_{\text{specular shading}}. \quad (3)$$



(a) Uniform Sampling (b) Importance Sampling

Fig. 3: Visualization of importance-weighted irradiance. The orange curve shows real-world illumination in general cases, where the sampled incident radiances $L_i(\omega_{ik})$ vary depending on each incoming direction. The blue (a) and green (b) curves represent the average of N sampled radiances, indicating the constant energy in sampled directions. By assuming that this constant radiance comes into a surface point p, the irradiance term can be separated from the integral in the rendering equation as shown in Equation 5. According to the surface roughness α, importance sampling selects optimal incident light directions for high-quality approximation. For example, when the surface is perfectly rough ($\alpha = 1$), lights are uniformly sampled (a) while sampled radiances only cover important specular directions to the viewing vector ω_o and the surface normal n, when α is in between 0 and 1 (b).

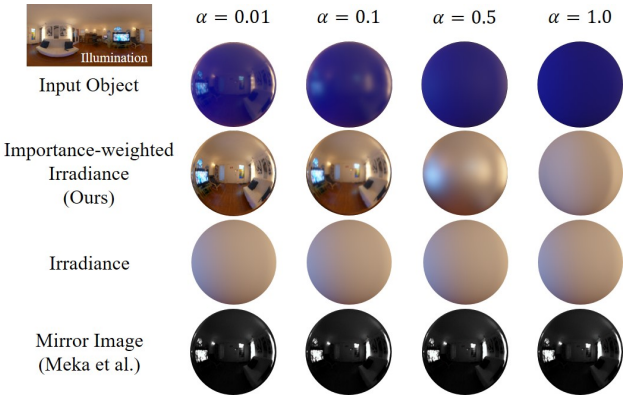


Fig. 4: Different types of irradiance images. Proposed importance-weighted irradiance images represent appropriate information depending on the surface roughness α. Irradiance images without importance sampling provide only low-frequency lighting information regardless of α. Mirror image [36] extracts mirror-reflected illumination in gray scale, which has difficulty in consistently estimating high-frequency information, especially from diffuse and rough surfaces like $\alpha = 1.0$.

3.2 Importance-weighted Irradiance

Inferring the information of incident radiances from a single image of an observed object requires us to consider the remaining factors of the rendering equation (Equation 3). For this reason, previous works used prior knowledge such as the known object geometry [19, 34, 40] or a particular material [19], as mentioned in Sect. 2.

To tackle this ill-posed problem without such assumptions, our first network, *Irradiance Net*, is designed to extract the average of incident radiances based on the importance sampling on each surface pixel, which is called *Importance-weighted Irradiance*. The purpose of estimating this term first is to effectively handle the unknown material terms from the input object observed under the rendering equation. Thus, with the reduction of dependency on the material, in the next step, we can better focus on estimating illumination by mapping importance-weighted irradiance on each surface point into an environment map in the directional domain (Sect. 3.3).

Note that we do not focus on estimating material properties as it was the case in previous works [17, 36], but on inferring illumination as the final output. Keeping this in mind, it is suitable to decrease the effects of other factors, except incident lights, in the early stage. Overall, distilling other factors makes our method more robust to various types of materials and increases the accuracy of predicted illumination when compared to the direct light estimation from a raw input object or prior methods, as discussed in Sect. 4. Before explaining the architecture of the irradiance net, we first show that the importance-weighted irradiance can be separated from other factors in the rendering equation.

Split sum approximation. To factor out importance-weighted irradiance from the integral form of the rendering equation, the split sum approximation [2, 28] is employed. This technique was originally introduced to handle heavy calculations of specular shading so that it performs efficient and realistic rendering in real-time games.

It starts with the unbiased Monte Carlo (MC) estimation [11], which estimates the integral of specular shading in Equation 3 by the sum of N different reflected radiances:

$$\int_{\Omega^+} \frac{DFG}{4(\omega_o \cdot n)(\omega_i \cdot n)} L_i(\omega_i)(\omega_i \cdot n) d\omega_i \approx \frac{1}{N} \sum_{k=1}^{N} \frac{DFG}{4(\omega_o \cdot n)(\omega_{ik} \cdot n)} \cdot \frac{L_i(\omega_{ik})(\omega_{ik} \cdot n)}{pdf(h_k, n, \alpha, \omega_o)}, \quad (4)$$

where importance sampling and the probability distribution function $pdf(h, n, \alpha, \omega_o) = D(h, n, \alpha)(h \cdot n)/4(\omega_o \cdot h)$ are used for generating sample lights based on half vector h, normal vector n, surface roughness α, and viewing vector ω_o, as proposed by Karis [28]. In this equation, it can be easily seen that incident light $L_i(\omega_i)$ is still multiplied with many other terms for computing the outgoing radiance value that is observed through each pixel.

In general cases, real-world illumination (the orange curve in Fig. 3) can be complex lighting with high frequencies. As a result, to compute the outgoing radiance, all incident radiances need to be considered together with material properties, the normal vector, and the viewing vector. On the other hand, the adopted approximation uses the mean of sampled incident radiances such as the blue curve of Fig. 3-(a) or the green curve of (b), and this mean radiance is considered with other terms. In other words, it is assumed that constant mean radiance comes into position p from every sampled solid angle of the hemisphere.

Based on this split sum approximation, the MC estimation to the specular shading is reformulated as follows:

$$\frac{1}{N} \sum_{k=1}^{N} \frac{DFG}{4(\omega_o \cdot n)(\omega_{ik} \cdot n)} \cdot \frac{L_i(\omega_{ik})(\omega_{ik} \cdot n)}{pdf(h_k, n, \alpha, \omega_o)} \approx \underbrace{\left(\frac{1}{N} \sum_{k=1}^{N} L_i(\omega_{ik}) \right)}_{\text{I-Irradiance}} \left(\frac{1}{N} \sum_{k=1}^{N} \frac{DFG}{4(\omega_o \cdot n) pdf(h_k, n, \alpha, \omega_o)} \right), \quad (5)$$

where the first term of the approximation, importance-weighted irradiance (I-Irradiance), is the average of all sampled incident radiances based on importance sampling, inherently depending on surface roughness α. To compose our ground truth dataset, such irradiance is rendered as RGB colors on each surface pixel. Basically, I-Irradiance is estimated from the specular shading of Equation 3, so in case of observing a diffuse material that has low specular shading, it may seem hard to extract such information. However, when the material becomes diffuse, incident radiances in the specular shading are sampled uniformly, which means that the I-Irradiance gets closer to the irradiance from the diffuse shading. As a result, we can also obtain I-Irradiance from the diffuse shading part, and our method can consistently estimate such adaptive irradiance regardless of specular and diffuse materials.

Note that importance sampling plays the key role in the split sum approximation because I-Irradiance can represent proper incident light information depending on α, making the rendering result of this approximation closer to that of the original rendering equation, as explained in Fig. 3-(b). Even though this estimation is biased, its main advantages are that it can effectively extract the I-Irradiance and decrease the complexity of unknown material properties, thus leading to a stronger focus on better light estimation.

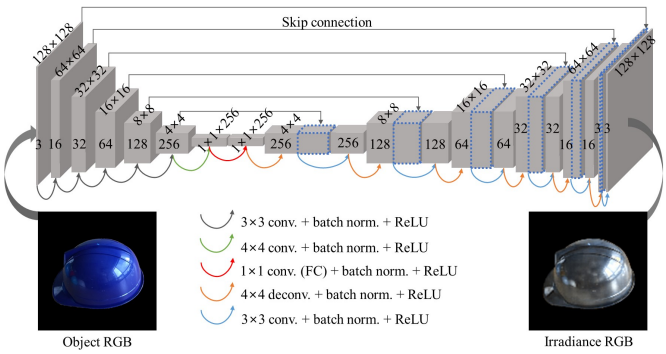


Fig. 5: The architecture of the irradiance net that estimates the importance-weighted irradiance from an input reference object.

To see the benefits of the I-Irradiance suitable for various input materials, examples with different roughness values given in Equation 5 are shown in Fig. 4, representing incident lights information from sharp to blurred. By using them as ground truth data in training the irradiance net, we can estimate high-frequency irradiance information on smooth specular surfaces, e.g., an input object with $\alpha = 0.01$, whereas more convoluted low-frequency irradiance can be inferred on rough surfaces, e.g., an input object with $\alpha = 1.0$. On the other hand, irradiance images based on uniform sampling, which have identical low-frequency incident light information regardless of surface roughness, provide very limited clues for light estimation. In case of the mirror image introduced by Meka et al. [36], a common point with ours is representing purer illumination reflected on object surfaces. However, our method conserves light colors in RGB channels, while the mirror image is in grayscale (Fig. 4) and it loses the important information of colored light sources. Furthermore, ours finds optimal irradiance according to the material's roughness, whereas the mirror image mainly aims to capture mirror reflection, causing difficulty when dealing with diffuse materials that do not contain sharp reflections. Those two approaches lead to different results of estimated illumination, and ours outperforms the previous one, as shown in Sect. 4.

The architecture of the irradiance net is described in Fig. 5. For high-quality estimation of I-Irradiance, U-Net [41] is employed with skip connections [20] that help in generating more detailed output images by decreasing the residuals between the input and expected output features. For a loss function, an L2 distance between ground truth and estimated I-Irradiance is used. More details of the network can be found in our supplementary material.

3.3 Estimation for LDR Illumination

To estimate LDR environmental illumination from the previously inferred importance-weighted irradiance image, a data-driven learning method is employed again. Since the dependency on the material was first decreased through the irradiance net, we can focus on finding an effective way to estimate radiances in each incident direction from the irradiance information on the input object. Each radiance with its own direction corresponds to a pixel with RGB values on an environment map that follows two-dimensional spherical coordinates. Thus, it requires domain transformation from the irradiance information on 2D surface pixels to respective incident lights in the directional domain.

Regarding this issue, previous methods directly used shape information [36] or assumed an observed object from a known class [39]. Since the proposed method aims to estimate illumination from a single RGB image without additional shape information such as depth, a normal map, or even a known class of an input object, such goal is extremely challenging. To tackle this issue, based on our large synthetic dataset covering various shapes and lighting conditions, the second network called *Illumination Net* learns a mapping function from importance-weighted irradiance on the object to illumination on an equirectangular environment map while decreasing dependency on the shape.

Furthermore, gradual training steps are adopted for more robust learning. As shown in Fig. 6, directly learning high-frequency illumination causes unstable loss results (blue), and a possible reason is that estimated high radiance values (e.g., light sources) slightly off the ground truth positions can result in an unnecessarily large loss with loss fluctuation, even though they are acceptable for IBL. Hence, to mitigate this issue, the second network is trained with convoluted illumination images, which gradually sharpen with increasing epochs, based on interpolation filter I [15]:

$$I(E,i,e) = \frac{1}{W}\sum_{k=1}^{N} E(\omega_k)(\omega_k \cdot n_i)^{\beta e}, \tag{6}$$

where E is an environment map on the sphere and n_i is a normal vector that starts from the center position of the sphere to the current pixel i to be convoluted. We choose N sample vectors ω_k around the normal vector, starting from the center position. Finally, sampled pixels $E(\omega_k)$ on the environment map are convoluted according to the weight $(\omega_k \cdot n_i)^{\beta e}$ and divided by the sum of weights $W = \sum_{k=1}^{N}(\omega_k \cdot n_i)^{\beta e}$. Here, β is the sharpness term, and e is the current epoch. During training, the filtered environment map changes gradually from the blurred image to the original high-frequency image. Filtered images are shown in Fig. 7. Here, e^2 is used as the sharpness term β. As the training steps go through 11 epochs, β changes from 1 to 121, so βe changes from 1 to 1331. Finally, learning on filtered images shows gradually better loss results than training on only high-frequency illumination, as shown in Fig. 6.

Through this regression, the network can infer an LDR distant light map, while each incident light gradually finds its right pixel position on the output environment map to decrease the total loss. It means that the direction of each light can be adjusted, completing the domain transformation. Although the illumination net does not directly use the geometry information of the reference object, it convincingly estimates LDR illumination because of our large dataset and gradual training method, as shown in Sect. 4. The architecture of the illumination net that uses the L2 loss function is described in Fig. 8, and more details can be found in our supplementary material.

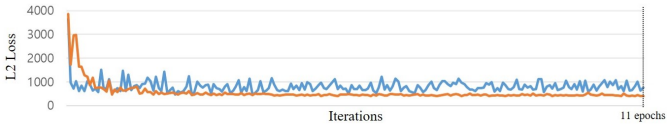


Fig. 6: Loss comparison between gradual and non-gradual trainings. When trained with high-frequency environment maps, loss results (blue) show larger amplitudes of vibration and a higher average loss than those of gradual training losses (orange). In the early stage of this graph, the gradual training is worse because blurred environment maps used in these steps lose much details of original high-frequency values.

Fig. 7: Filtered LDR environment maps according to the changing epochs. From the top-left to the bottom-right, βe changes from 1 to 1000, gradually providing sharper and higher frequency images.

3.4 Reconstruction for HDR Illumination

Since image-based lighting (IBL) [10] requires HDR radiance values to render photo-realistic virtual objects coherently with the real-world environment, the LDR illumination estimated through the illumination net is still insufficient. Thus, as a final step, an HDR environment map is reconstructed by combining estimated LDR illumination and HDR radiance values inferred from the third network called *HDR Net*.

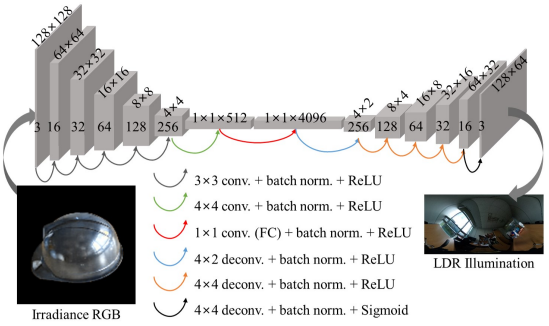


Fig. 8: The architecture of the illumination net performing domain transformation from object surfaces to an LDR environment map.

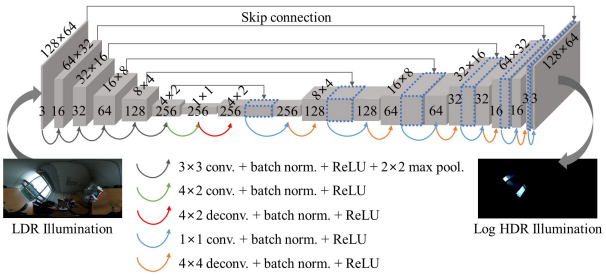


Fig. 9: The architecture of the HDR net estimating HDR radiance values in the saturated areas. An input image is a previously inferred LDR environment map through the illumination net and an output contains HDR radiance values in the log space.

To accurately recover a camera response function from a single LDR input image, it is necessary to estimate related camera parameters such as exposure, ISO, and white balance values. Since it is found that learning those parameters from limited input information is problematic, resulting in a high variance in estimation [12], the proposed method does not concentrate on estimating a whole camera response function, but on recovering HDR values only from saturated areas such as light sources, which is the key factor for improving rendering quality in IBL.

When directly learning linear HDR values, even small pixel areas with high radiance values can have a significant effect on the estimation of unsaturated areas. To alleviate this problem, the HDR net is trained in the log domain (log_{10}(HDR radiance)), and $exp(H_{i,c})$ is used for final reconstruction. Let $H_{i,c}$ denote a result of the HDR net given pixel index i and RGB channel c.

Once LDR illumination estimated from the illumination net and recovered HDR values through the HDR net are obtained, those two types of radiances are combined as final HDR values denoted as $R_{i,c}$, to be used in IBL, as follows [12]:

$$R_{i,c} = (1-\gamma_i) f^{-1}(U_{i,c}) + \gamma_i exp(H_{i,c}), \tag{7}$$

where for each RGB color channel c of a pixel index i, the reconstructed HDR radiance $R_{i,c}$ is linearly interpolated between two estimated HDR values $f^{-1}(U_{i,c})$ and $exp(H_{i,c})$. $U_{i,c}$ is an estimated LDR value from the illumination net. The blending weight γ_i is defined as

$$\gamma_i = \frac{max(0, max_c(U_{i,c}) - \tau)}{1-\tau}, \tag{8}$$

where $\tau = 0.8$ when input $U_{i,c}$ has a range of $[0,1]$. This blending method means that HDR values in unsaturated areas can be recovered by using estimated LDR inputs and the inverse camera response function f^{-1}, assuming that the response functions of casual cameras used in MR applications can be approximately represented by the homogeneous sigmoid function [12], and the inverse sigmoid function is f^{-1}. On the other hand, in saturated areas, the estimated HDR value from the HDR net is used for more accurate reconstruction of HDR radiance. The architecture of the HDR net is described in Fig. 9, and more details can be found in our supplementary material. As a loss function, the L2 distance between an estimated log HDR environment map and its ground truth image is used.

3.5 Learning Details and Dataset

All networks are trained by using Caffe [23] on two GTX 1080Ti GPUs. Adam optimizer [29] is used with momentum parameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ as generally used. The fixed learning rate is 0.001, and weight decay is 0.0001, related to regularization. Until showing the smallest test losses, the irradiance net, illumination net, and HDR net take 20 (6 epochs), 24 (11 epochs), and 14 (124 epochs) hours, respectively, with a batch size of 16.

Training on a large quantity of realistic data can significantly improve the estimation quality of deep neural networks. Hence, many images of reference objects based on the physically correct rendering equation (Equation 3) are prepared synthetically. This effectively substitutes a real dataset requiring laborious works to obtain. In our case, we use real images to test feasibility of our method in real-world applications.

A wide variety of images of reference objects are synthesized with different conditions such as shapes, camera viewpoints, types of illumination, and materials. 403 indoor and outdoor HDR images are obtained from websites [5, 46] and the database of Gardner et al. [15]. With Ricoh Theta V, outdoor HDR images are also captured at 147 different spots in a university campus. High-resolution images of 5357×2688 are taken with different exposure values from EV -3.0 to +3.0, followed by the synthesizing step for an HDR image in Photoshop CC 2017. Among a total of 550 HDR images, 50 images are used for tests. In addition, 35 (26 for training and 9 for testing) different 3D models are downloaded from a free website [14], including cars, animals, and so on. When choosing them, concave shapes are avoided because our method does not consider occluded incident lights. In terms of materials, as the rendering algorithm is based on the Cook-Torrance BRDF (Equation 2), images are rendered based on varying values for material parameters, including IOR, roughness, and diffuse albedo. Specifically, those material properties are randomly selected for each rendering, resulting in 8 random materials, including metal and non-metal from diffuse to specular, with well balanced quantities. Finally, about 2.3 million training images are rendered for training. These are synthetically constructed by placing 26 different objects at the center of the hemisphere, with 500 environment maps and 8 random materials at 23 different camera viewpoints evenly distributed over the upper hemisphere. Also, their importance-weighted irradiance images are rendered based on Equation 5.

To render the input LDR training dataset for the HDR net, we use a parametric function [12] fitted to simulate camera response functions collected by Grossberg and Nayar [18], with randomly selected exposure values ranging from 0.5 to 2.0. LDR environment maps are simulated from 500 HDR images by changing 23 different camera

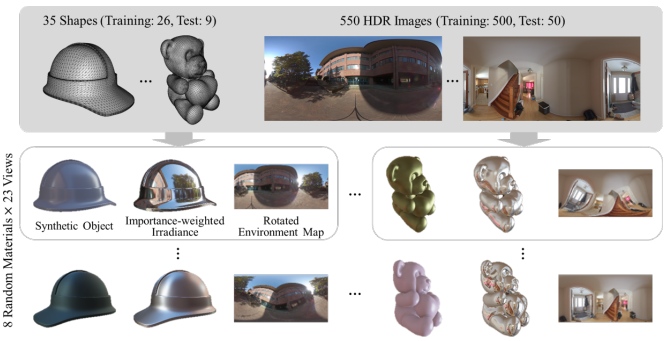


Fig. 10: The synthetic training dataset rendered by our own renderer. Objects have various shapes, illuminations, and materials. Overall, our synthetic dataset has 2.3 million image pairs.

Table 1: Quantitative comparison of our network designs. Compared with *Ours*, *Direct* uses a raw input object to directly estimate illumination through the illumination net without inferring I-Irradiance and gradual training. *Irradiance* infers I-Irradiance in the first step but estimates illumination without gradual training. *Ideal* indicates the upper bound that each network can achieve when using a ground truth input.

	I-Irradiance			LDR Illumination			HDR Illumination		
	PSNR	RMSE	SSIM	PSNR	RMSE	SSIM	PSNR	RMSE	SSIM
Direct	-	-	-	13.9374	0.2117	0.3740	25.5962	0.0577	0.3701
Irradiance	26.9803	0.0449	0.9502	16.4243	0.1588	0.4459	29.5844	0.0409	0.4284
Ours: I-Irradiance + gradual	26.9803	0.0449	0.9502	**17.8786**	**0.1271**	**0.5298**	**34.8689**	**0.0181**	**0.5594**
Ideal: I-Irradiance + gradual	26.9803	0.0449	0.9502	18.8466	0.1148	0.5745	36.1276	0.0156	0.9383

views and randomly selecting 5 exposure values for each rendering in order to obtain 57,500 images for training. All of the synthetic images are rendered on our own renderer, as shown in Fig. 10.

4 Results and Analysis

4.1 Evaluation on Network Performance

Here, our selected design is compared to the other two designs in order to validate the necessity of the proposed intermediate steps for inferring I-Irradiance and the gradual training technique employed. *Direct* in Table 1 estimates LDR illumination from a raw input object without the irradiance net. *Irradiance* uses the irradiance net so that it estimates LDR illumination from the I-Irradiance image. However, it is trained to directly learn high-frequency illumination, which shows unstable loss convergence, as shown in Fig. 6. *Ours (I-Irradiance + gradual)* uses the I-Irradiance image to decrease material dependency first and then infers LDR illumination by using the illumination net that gradually learns from low- to high-frequency illumination. In *Ideal*, each network uses a ground truth input instead of an estimated result from the previous step, achieving the best performance of each network.

For each network, errors between estimations and ground truth images are calculated by using three different metrics, including PSNR, RMSE, and SSIM, and all of the experiments are performed by using our synthetic test dataset that was never shown in the training step. From the dataset, 1k image pairs were randomly chosen, including synthetic input objects rendered with 9 different shapes, random materials, and 50 environments. We keep balance not only between specular and diffuse materials, but also between indoor and outdoor environments. In this work, the roughness $\alpha = 0.5$ is used to separate materials into specular and diffuse ones.

I-Irradiance results show the estimation quality of importance-weighted irradiance. Specifically, RMSE shows a low mean error (0.0449) when RGB pixel values range from 0 to 1 and are measured in a whole image with 128×128 resolutions. It is found that employed skip connections effectively enhance the accuracy of the estimated irradiance by learning to reduce the residual errors between the input and output features in the same domain, so that the irradiance net can well decrease the dependency on the input object's material. Fig. 11 shows visually reliable irradiance estimated using the proposed method.

Meaningful differences come out from the results of estimated LDR illumination. Here, since the learned domain transformation is performed in the illumination net, estimating exact pixel intensities in their right positions is highly challenging. Thus, in terms of estimating affordable illumination for IBL, SSIM can be a better metric to assess LDR illumination based on the similarity of the structure, contrast, and luminance. In the obtained result, compared to *Direct* design, *Irradiance* achieves 19% improvement in terms of SSIM, proving the advantage of using the irradiance net that provides only irradiance information as input to the next step. Moreover, when gradual training steps are employed in the proposed network design (*I-Irradiance + gradual*), another 18% higher SSIM quality of estimated LDR illumination is achieved. This gradual technique therefore plays an important role in learning this challenging domain transforming problem.

In the results of HDR illumination, *Direct* and *Irradiance* approaches show lower SSIM results than those of their own LDR estimations. By contrast, the proposed approach achieves better SSIM of HDR estimation than that of LDR illumination. We think that this interesting result comes from the fact that *Ours* estimates more accurate light positions in LDR illumination, so HDR radiances in saturated areas can be restored well, thus improving SSIM quality. Note that a predicted HDR image is tone-mapped to the RGB domain when evaluated in SSIM, and PSNR and RMSE calculate errors with log HDR radiances, as adopted in the previous work [38]. For intuitive illustration, final HDR values are visualized as a heat map that ranges from 0 to 10, as shown in Fig. 11. Although there is a noticeable gap of SSIM between *Ours* and *Ideal* results due to the structural errors from the challenging domain transformation in the earlier step, RMSE shows only a small difference, indicating that our method successfully reconstructs HDR information, especially in saturated areas like light sources.

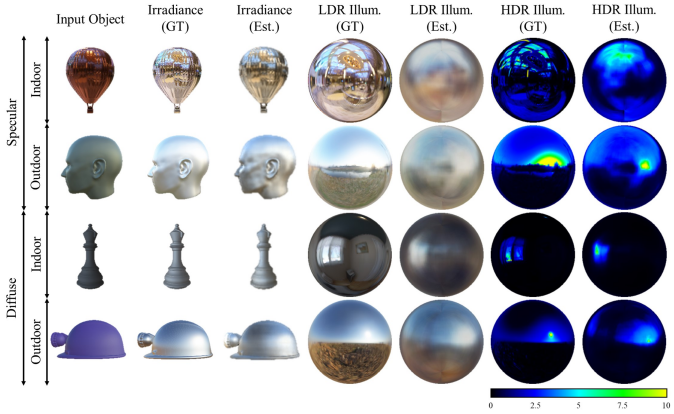


Fig. 11: Results of proposed sequential networks using the synthetic test dataset. For testing in various conditions, we use input objects with arbitrary specular and diffuse materials in indoor and outdoor environments. Here, the terms *Illum.*, *GT* and *Est.* denote illumination, ground truth, and estimation, respectively. Reconstructed HDR radiance is represented by a heat map where the value ranges from 0 to 10.

4.2 Comparisons

Illumination comparisons. Here, the proposed method is compared to state-of-the-art approaches by using our synthetic test dataset first and then a real dataset for validating the applicability of the proposed method in MR. For the real dataset, 80 images of 15 real objects were captured, including various shapes, materials, and camera views in indoor and outdoor scenes. Especially, in all synthetic and real test images, a balance between specular and diffuse materials of input objects was kept to cover the cases of observing various types of materials.

As comparable recent methods, the results of Georgoulis et al. [38] and Meka et al. [36] are used. Note that the direct method of Georgoulis et al. is used to infer the illumination without normals, same as ours, and object normals are necessary for Meka et al. to convert a mirror image into an environment map. For a fair comparison, all those methods are trained on our synthetic dataset, and their results are represented by a reflectance map. Because estimation of Meka et al. does not provide HDR information, tone-mapping is performed for the results of the others, and the estimation is compared with ground truth LDR illumination that also uses the same tone-mapping method.

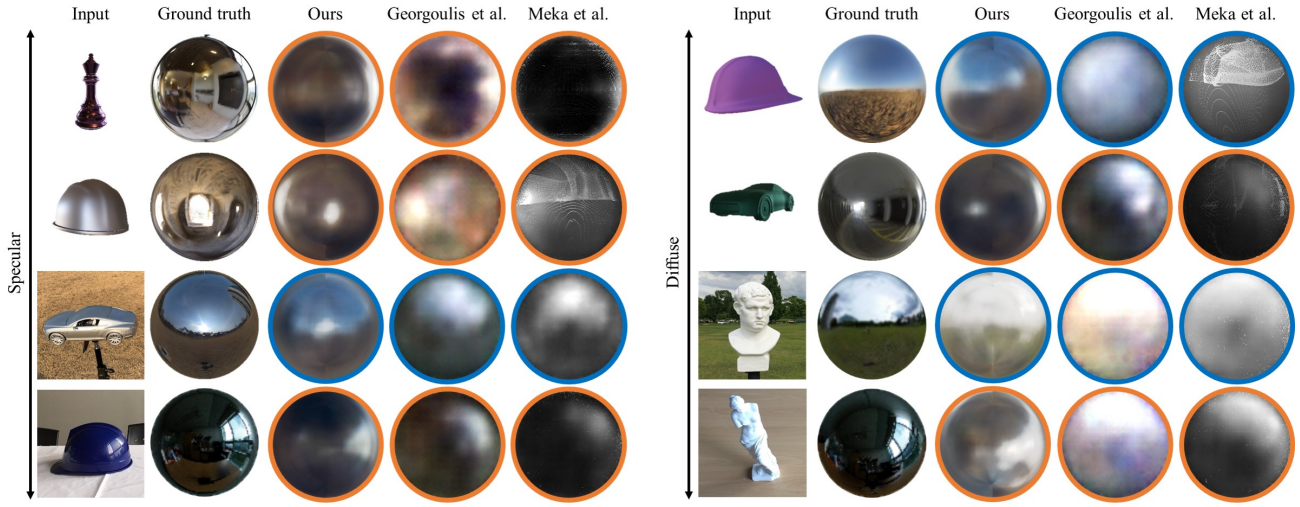


Fig. 12: Qualitative comparison with previous approaches using our synthetic and real test dataset. The first two rows use synthetic input images while bottom two rows estimate illumination from real objects. Furthermore, left and right parts are divided based on the different material types of input objects. Note that the orange and blue circles represent estimated indoor and outdoor illumination respectively.

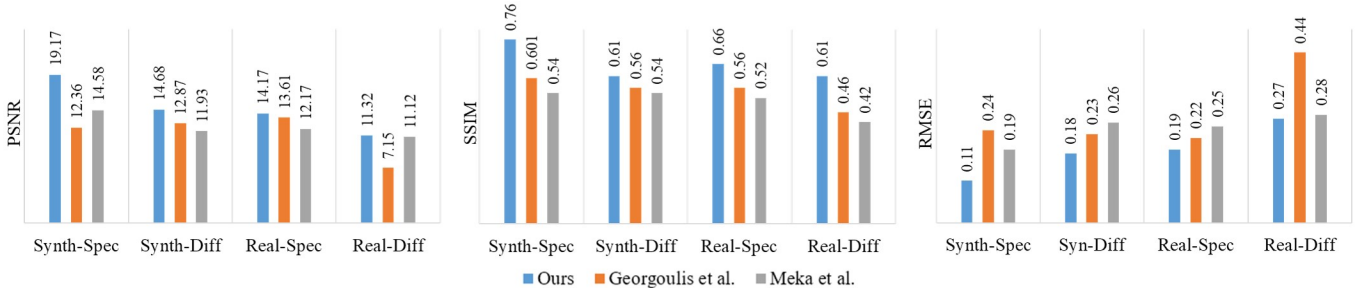


Fig. 13: Quantitative comparisons. For each metric, four types of test sets are evaluated, including synthetic-specular (Synth-Spec), synthetic-diffuse (Synth-Diff), real-specular (Real-Spec), and real-diffuse (Real-Diff). Higher values are better for PSNR and SSIM, except RMSE.

For a qualitative comparison, the results of three different methods are visualized in Fig. 12. The proposed method visually estimates illumination better in terms of incident light directions and RGB colors. For example, in the first row of the right part, ours can catch relatively correct light directions and colors when compared with the other methods. Even though Meka et al. uses an additional normal map to estimate higher-frequency illumination, it only provides gray-scaled intensities. However, ours infers colors of white sunlight, sky, and the brown ground well. This method can also predict the geometrical structure of an environment map as seen in the estimation results of the bottom row that estimate warped widows, so they provide a more accurate direction of incoming lights. In particular, when an input object consists of a diffuse material, which is a more challenging condition, ours provides qualitatively better estimation than the others.

In the quantitative results in Fig. 13, our method outperforms other two methods in terms of PSNR, SSIM, and RMSE across all test conditions, including synthetic and real objects with specular and diffuse materials. Although estimating illumination from a real-world object is a more challenging task for the proposed method because our training dataset is totally composed of synthetic images, there is a reasonably small gap between synthetic and real test results. Specifically, average SSIM results from synthetic and real test images are about 0.68 and 0.63, respectively. Note that SSIM represents 1.0 when the two compared images are identical. We think that the main reason of this slight difference comes from our physically-motivated network design and photo-realistic training images that use the physically-based BRDF.

One interesting finding is that our method gets the highest improvement over other works when observing a real-world diffuse object. We achieve about 30% and 40% quality enhancements when compared with the SSIM results of Georgoulis et al. and Meka et al., respectively, proving that our strength is shown with diffuse materials. The main reason for this could be the fact that our method extracts proper irradiance information even from a diffuse material through the irradiance net, which is a remarkable difference compared to other methods.

Relighting comparisons. We also provide additional comparisons in terms of relighting quality. Since coherently relit objects help users get immersed in MR, these comparison results may have important meaning. We first render virtual objects with ground truth HDR environment maps included in our real testing dataset and compare them to the relit virtual objects rendered with estimated illumination from three different works. In Fig. 14, there are four scenes, and each image has a real object (left) and a relit virtual object (right). Please focus on the differences in the estimated geometry structures reflected on the plastic hats in the first row, shiny effects in the second row, incident lights from the right in the third row, and reflected highlights at the back of the bunnies in the last row. Our method not only qualitatively provides better results, but also has the lowest RMSE results among the three state-of-the-art works, as shown in the upper right part of each image.

5 Discussion and Conclusion

This paper presented a novel deep learning-based method for estimating HDR illumination from a single RGB image of an observed object with a homogeneous material and mostly convex shape. Our method uses three end-to-end networks based on physically-motivated designs. Through sequentially decreasing dependency on material and shape, high-quality HDR illumination of an indoor or outdoor scene can be effectively estimated when there is a reference object. Experiment results showed that our method outperformed state-of-the-art approaches, espe-

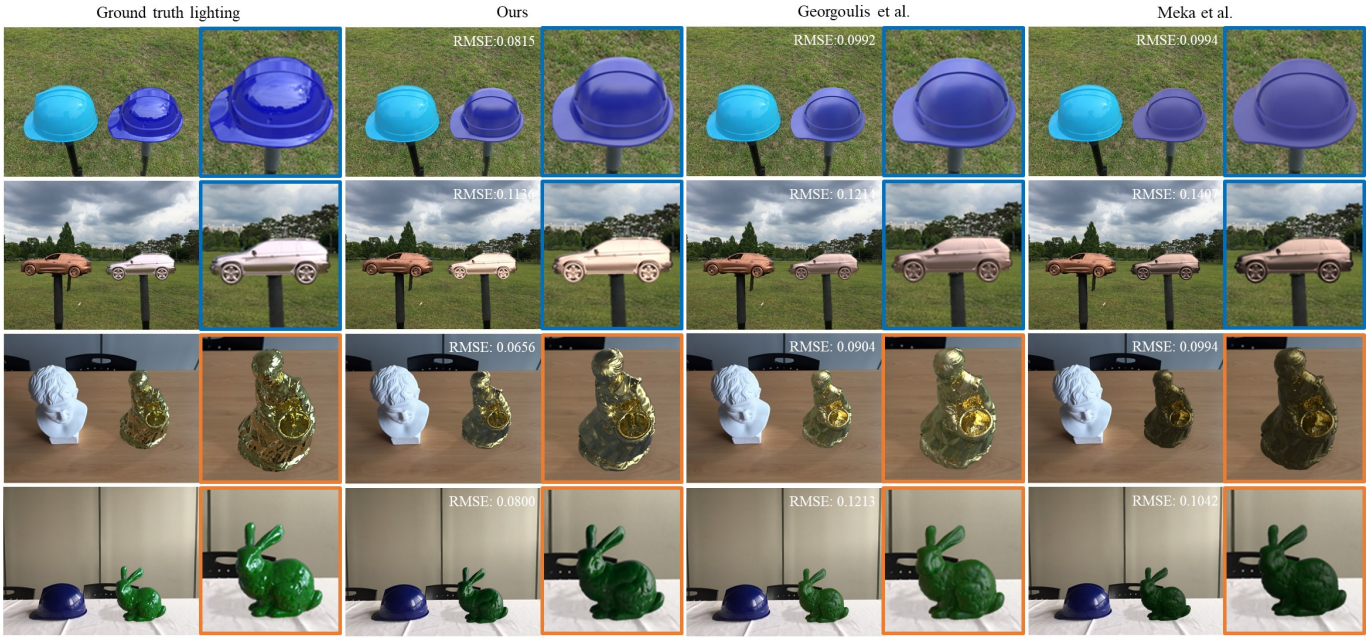


Fig. 14: Comparison on our real test images. In each image, the individual method estimates the illumination by observing a real-world object on the left side. Then, with the predicted illumination, we relight a virtual object next to the real object. An RMSE value between the ground truth and the relit object is at the upper right corner. Please refer to the corresponding part of the paper for more detailed descriptions.

cially in terms of covering a wide range of materials, including specular and diffuse ones. To underpin the applicability of this method in MR or image synthesis, some interesting examples are shown in Fig. 15. Furthermore, in the supplementary video, we show the availability of our method in real-time MR that involves dynamically changing illumination. Regarding the performance, when an Intel Core i7-6700k CPU is used, average execution times of the sequential networks are 24.37 ms, 19.68 ms, and 12.81 ms, in order. In case of using an NVIDIA GeForce GTX 1080Ti GPU, respective networks show even faster performances, which are 9.68 ms, 8.12 ms, 3.44 ms, in sequence.

Nonetheless, our method has certain limitations to overcome. Because the employed lighting model and synthetic dataset only consider reflection on a homogeneous material, the proposed networks cannot cover more complex materials such as spatially varying or translucent ones. As an example shown in Fig. 16, our method imperfectly estimates importance-weighted irradiance when an input has a heterogeneous material or texture, showing a tendency to only decrease a dominant homogeneous material. In addition, even though using log scaled HDR radiance is beneficial in training the proposed HDR net, it is still hard to estimate accurate and extremely high-frequency HDR illumination, which should be addressed in our future work. Although the proposed learning-based method can estimate illumination from unseen directions, it is also affected by incomplete information of an observed object which has high occlusion or insufficient normal vectors. Finally, our real-time system shown in the video also needs advanced techniques, including temporally coherent light estimation considering sequential frames when learning. In addition, this system uses an object segmentation method based on depth cutting and it provides some noisy inputs to the irradiance net, causing flickering effects and estimation errors. For those issues, learning-based segmentation methods [8, 36] can be employed, which are robust to general and dynamic backgrounds.

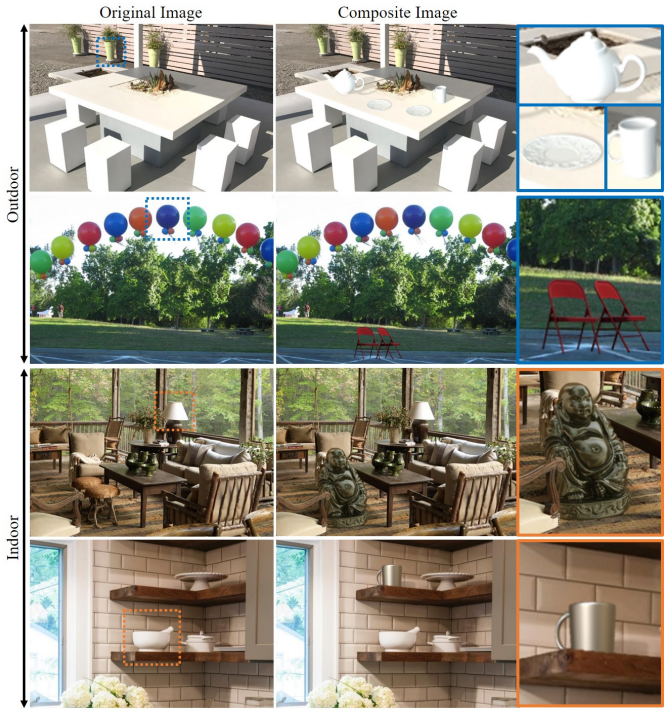


Fig. 15: MR applications. We select random images from the internet, and estimate the HDR illumination from the objects. Observed indoor and outdoor objects are marked by orange and blue dotted boxes respectively. Using only a single RGB image, our method provides a realistic MR scene by inserting coherently rendered virtual objects. Images are from archiexpo, balloondecorofcentralcalifornia, hgtv, homedit.

Fig. 16: Failure cases of the proposed method when observing heterogeneous material-objects. Especially, when an object has a complex texture (the right object), the irradiance net imperfectly decreases dependency on materials, resulting in low-quality illumination estimation.

ACKNOWLEDGMENTS

This work was partly supported by Institute for Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2019-0-01648, Development of 360 degree VR content authoring platform based on global street view and spatial information) and Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT (NRF-2017M3C4A7066316).

REFERENCES

[1] M. Aittala. Inverse lighting and photorealistic rendering for augmented reality. *The Visual Computer*, 26(6-8):669–678, 2010.
[2] T. Akenine-Moller, E. Haines, and N. Hoffman. *Real-time rendering*. AK Peters/CRC Press, 2018.
[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*, pp. 17–24, 2013.
[4] J. F. Blinn. Models of light reflection for computer synthesized pictures. In *ACM SIGGRAPH computer graphics*, vol. 11, pp. 192–198. ACM, 1977.
[5] C. Bloch. Hdr labs. http://www.hdrlabs.com.
[6] 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*, vol. 37, pp. 51–61. Wiley Online Library, 2018.
[7] L.-C. Chen, X.-L. Nguyen, and C.-W. Liang. Object segmentation method using depth slicing and region growing algorithms. In *International Conference on 3D Systems and Applications, Tokyo, Japan*, pp. 87–90, 2010.
[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
[9] R. L. Cook and K. E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics (TOG)*, 1(1):7–24, 1982.
[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*, pp. 189–198. ACM, 1998.
[11] A. Doucet, N. De Freitas, and N. Gordon. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pp. 3–14. Springer, 2001.
[12] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM Transactions on Graphics (TOG)*, 36(6):178, 2017.
[13] E. Fernandez-Sanchez, J. Diaz, and E. Ros. Background subtraction based on color and depth using active sensors. *Sensors*, 13(7):8895–8915, 2013.
[14] Free3D. Free3d. https://www.free3d.com.
[15] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (TOG)*, 36(6):176, 2017.
[16] S. Georgoulis, K. Rematas, T. Ritschel, M. Fritz, T. Tuytelaars, and L. Van Gool. What is around the camera? In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5170–5178, 2017.
[17] 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.
[18] M. D. Grossberg and S. K. Nayar. What is the space of camera response functions? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. II–602, 2003.
[19] L. Gruber, T. Richter-Trummer, and D. Schmalstieg. Real-time photometric registration from arbitrary geometry. In *IEEE international symposium on mixed and augmented reality (ISMAR)*, pp. 119–128, 2012.
[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
[21] 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*, vol. 2, 2017.
[22] B. K. Horn and R. W. Sjoberg. Calculating the reflectance map. *Applied optics*, 18(11):1770–1779, 1979.
[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
[24] J. T. Kajiya. The rendering equation. In *ACM Siggraph Computer Graphics*, vol. 20, pp. 143–150. ACM, 1986.
[25] P. Kán. Interactive hdr environment map capturing on mobile devices. In *Eurographics (Short Papers)*, pp. 29–32, 2015.
[26] P. Kán and H. Kafumann. Deeplight: light source estimation for augmented reality using deep learning. *The Visual Computer*, 35(6-8):873–883, 2019.
[27] P. Kán and H. Kaufmann. High-quality reflections, refractions, and caustics in augmented reality and their contribution to visual coherence. In *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 99–108. IEEE, 2012.
[28] B. Karis. Real shading in unreal engine 4. *SIGGRAPH Physically Based Shading Theory Practice course*, pp. 621–635, 2013.
[29] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
[30] M. Knecht, C. Traxler, O. Mattausch, W. Purgathofer, and M. Wimmer. Differential instant radiosity for mixed reality. In *2010 IEEE International Symposium on Mixed and Augmented Reality*, pp. 99–107. IEEE, 2010.
[31] S. B. Knorr and D. Kurz. Real-time illumination estimation from faces for coherent rendering. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 113–122. IEEE, 2014.
[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.
[33] S. Liu and M. N. Do. Inverse rendering and relighting from multiple color plus depth images. *IEEE Transactions on Image Processing*, 26(10):4951–4961, 2017.
[34] 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.
[35] D. Mandl, K. M. Yi, P. M. Mohr, P. M. Roth, P. Fua, V. Lepetit, D. Schmalstieg, and D. Kalkofen. Learning lightprobes for mixed reality illumination. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 82–89. IEEE, 2017.
[36] A. Meka, M. Maximov, M. Zollhöfer, A. Chatterjee, H.-P. Seidel, C. Richardt, and C. Theobalt. Lime: Live intrinsic material estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6315–6324, 2018.
[37] S. Pessoa, G. Moura, J. Lima, V. Teichrieb, and J. Kelner. Photorealistic rendering for augmented reality: A global illumination and brdf solution. In *Virtual Reality Conference (VR), 2010 IEEE*, pp. 3–10. IEEE, 2010.
[38] K. Rematas, S. Georgoulis, T. Ritschel, E. Gavves, M. Fritz, L. Van Gool, and T. Tuytelaars. Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE transactions on pattern analysis and machine intelligence*, 2017.
[39] 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 (CVPR)*, pp. 4508–4516, 2016.
[40] T. Richter-Trummer, D. Kalkofen, J. Park, and D. Schmalstieg. Instant mixed reality lighting from casual scanning. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 27–36, 2016.
[41] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.
[42] T. Schwandt and W. Broll. A single camera image based approach for glossy reflections in mixed reality applications. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 37–43. IEEE, 2016.
[43] P.-P. Sloan, J. Kautz, and J. Snyder. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. In *ACM Transactions on Graphics (TOG)*, vol. 21, pp. 527–536. ACM, 2002.
[44] L. Xu, J. S. Ren, C. Liu, and J. Jia. Deep convolutional neural network for image deconvolution. In *Advances in Neural Information Processing Systems*, pp. 1790–1798, 2014.
[45] S. Yoon. *Rendering*. Freely available on the internet, 1 ed., 2018. https://sglab.kaist.ac.kr/ sungeui/render/.
[46] G. Zaal. Hdri haven. https://www.hdrihaven.com.
[47] M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *International Conference on Computer Vision*, pp. 2018–2025. IEEE, 2011.