Adaptive Light Estimation using Dynamic Filtering for Diverse Lighting Conditions

Junhong Zhao, Andrew Chalmers, and Taehyun Rhee

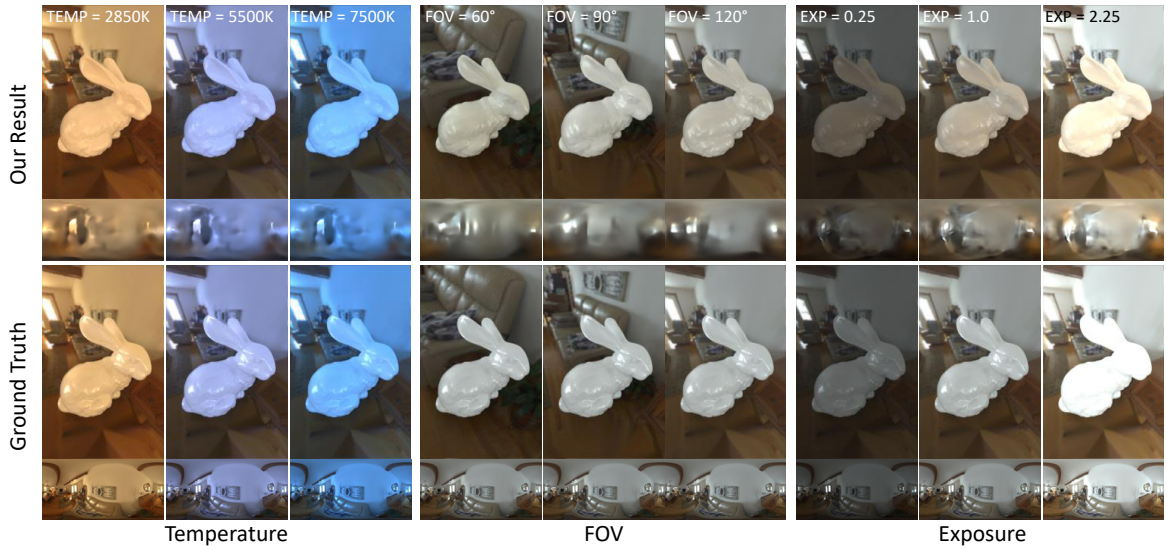


Fig. 1: Given a regular RGB photograph from the same indoor scene (background of the bunny) but with different color temperature (top), exposure (middle), and field-of-view (FOV, bottom), our method predicts the HDR environment map (below each photograph of a bunny; ground truth and our prediction). The Stanford Bunny is rendered and composited into the input photograph to show the illumination effect. We observe consistent lighting with varying color temperature and exposure. As FOV increases, we observe that the light estimate is stable and improves as more lighting information is visually present within the FOV.

Abstract— High dynamic range (HDR) panoramic environment maps are widely used to illuminate virtual objects to blend with real-world scenes. However, in common applications for augmented and mixed-reality (AR/MR), capturing 360° surroundings to obtain an HDR environment map is often not possible using consumer-level devices. We present a novel light estimation method to predict 360° HDR environment maps from a single photograph with a limited field-of-view (FOV). We introduce the Dynamic Lighting network (DLNet), a convolutional neural network that dynamically generates the convolution filters based on the input photograph sample to adaptively learn the lighting cues within each photograph. We propose novel Spherical Multi-Scale Dynamic (SMD) convolutional modules to dynamically generate sample-specific kernels for decoding features in the spherical domain to predict 360° environment maps. Using DLNet and data augmentations with respect to FOV, an exposure multiplier, and color temperature, our model shows the capability of estimating lighting under diverse input variations. Compared with prior work that fixes the network filters once trained, our method maintains lighting consistency across different exposure multipliers and color temperature, and maintains robust light estimation accuracy as FOV increases. The surrounding lighting information estimated by our method ensures coherent illumination of 3D objects blended with the input photograph, enabling high fidelity augmented and mixed reality supporting a wide range of environmental lighting conditions and device sensors.

Index Terms—Augmented reality, mixed reality, lighting, light estimation, deep learning

1 Introduction

Lighting is one of the key components to seamlessly blend virtual objects into a photograph for augmented and mixed reality (AR/MR). For high fidelity illumination coherent with the background image, the environmental lighting information beyond the limited field-of-view (FOV) of a typical photograph is required. A 360° omnidirectional capturing device or light probe can be used to obtain a panoramic, high dynamic range (HDR) environment map [8], but such setups are often not available for conventional AR/MR applications (e.g., mobile phone).

To address this problem, inverse rendering techniques can be employed to recover the HDR environment map from a single, standard, low dynamic range (LDR), limited FOV photograph. This is a challenging and often ill-posed problem [11, 23], as it requires estimating lighting properties dependent on the observable illumination cues (e.g., specular highlights, shadows and tone) within a photograph. Since the

- *Junhong Zhao is with Computational Media Innovation Centre (CMIC), Victoria University of Wellington, New Zealand. E-mail: junhong.jennifer@gmail.com.*
- *Andrew Chalmers is with Computational Media Innovation Centre (CMIC), Victoria University of Wellington, New Zealand. E-mail: andrew.chalmers@vuw.ac.nz.*
- *Taehyun Rhee is with Computational Media Innovation Centre (CMIC), Victoria University of Wellington, New Zealand. E-mail: taehyun.rhee@vuw.ac.nz.*

Manuscript received 15 Mar. 2021; revised 11 June 2021; accepted 2 July 2021. Date of publication 27 Aug. 2021; date of current version 1 Oct. 2021. Digital Object Identifier no. 10.1109/TVCG.2021.3106497

photograph itself only captures a limited FOV of the scene, estimating lighting information outside the FOV of the photograph is challenging as it depends on the visible illumination cues. There are two primary factors that affect illumination cues to consider: 1) the environmental factors and 2) the sensor/lens characteristics of the capturing device. The environmental factors may be composed of objects with varying appearance, luminaries with varying color temperatures and intensities, and illumination variations across indoor or outdoor environments. Whereas sensors/lens' can have varying FOV, apertures, and white balance settings. As such, inverse rendering techniques must be selective in identifying the lighting-specific features within the photograph, robust to the environmental and sensor factors, to provide high quality lighting estimation.

Recent methods have used deep learning to solve this problem, beginning with the seminal work by Gardner et al. [11] and Hold-Geoffroy et al. [17] to estimate indoor and outdoor lighting using a convolutional neural network (CNN) respectively. Since then different methods have addressed specific critical issues that arise from lighting estimation, such as accounting for spatially varying indoor lighting [12, 24, 35, 40], outdoor non-parametric lighting [43], and mobile mixed reality [4, 6, 23], among others. However, such CNN architectures, once trained, have no flexibility in dynamically adjusting to the lighting cues that are apparent within each individual photograph. This will limit the generalization of the CNN models to arbitrary scenes with high variations in lighting cues.

We propose an adaptive deep model, DLNet, that can adjust to the input photograph dynamically for environmental lighting estimation. Our key idea is to build and train a deep neural network that is aware of the wide variations of input photographs. Inspired by the idea of dynamic filtering [19], we introduce a spherical dynamic filtering framework that adaptively generates the convolution filters tailored for each input sample, as opposed to the traditional framework where the filters are fixed after training and shared among all samples. We impose novel spherical multi-scale dynamic (SMD) convolutional modules onto multiple layers of the decoder, and a multi-scale dynamic (MD) module onto the feature extraction layers. The dynamic features extracted from dynamic convolutions are more selective for lighting-specific cues and improve generalization of light estimation. Using a training dataset augmented with respect to FOV, an exposure multiplier (hereafter referred to as *'exposure'*), and color temperature, our proposed network estimates environmental lighting with higher accuracy than the prior work. We evaluated our method with different input conditions with varying FOV, exposure, and color temperature. Our estimation shows, unlike prior work, consistent light structure with variations of exposure and color temperature, and improved accuracy with increased FOV (Figure 1). Our light estimation is integrated into an AR/MR framework to provide high fidelity illumination.

2 Related Work

Recent methods use deep learning frameworks to estimate lighting from photographs, each addressing different sub-problems within the field of inverse rendering. To address this challenge, a physically inspired approach can be undertaken [13, 30], which uses a foreground object with unknown geometry or reflectance as the input that is then intrinsically decomposed [3] into properties that define its appearance (material, geometry, and lighting). Appearance maps [26], reflectance maps [13], or both illumination and reflectance [31] are recovered after the decomposition process with deep learning. Other methods directly relight the input images according to the new lighting conditions for arbitrary objects [32, 39] or faces [29, 37, 44].

Alternatively, when there is no clear exemplar object present within the photograph, the general lighting cues such as shadow, shading, highlights, and occlusion relationship need to be used instead to infer the lighting. Our work follows this setup, where we assume an unconstrained, flexible setting without any specific reference object provided. Further, prior work focus on either indoor [10, 11, 35, 36, 40, 41] and/or outdoor [10, 16, 21, 22] scenes. We choose to focus our experiments on indoor scenes to evaluate our method against the wide variation of color temperatures and intensities emitted from light bulbs.

Within the space of indoor light estimation, special consideration has been given to spatially varying nature of lights with a fixed position (as opposed to outdoor lighting, where the sun is considered effectively infinitely far away). This was first addressed through a spatially warping operator [11], but was extended by considering depth estimates of the photograph [35, 40] as part of the training procedure. Parallax provided by stereo pairs of photographs has also been used to address this issue [36]. In our work, we found the spatially warping operator to be fast and sufficient. We also avoid stereo pairs to provide wider support for consumer level devices with monocular cameras. Other recent work has been designed specifically for mobile applications either targeting light-weight mobile neural networks [23], or producing efficient lighting representations for rendering such as spherical harmonics [6] or reflection maps [4].

Although many neural network structures have been proposed to address various inverse lighting problems, special consideration has not been given to the wide range of variations of lighting found between the environment and camera response. Prior methods share the model parameters among different samples. As such, they do not adjust to the wide range of characteristics of each new input. In this work, we propose to adaptively learn lighting features from different samples inspired by the dynamic filtering method [19], which has been evidenced to be beneficial to adaptive feature extraction and implicit temporal/spatial transformations learning [14, 15, 20, 38]. We introduce a novel SMD convolutional module to enhance the generalization ability of the light estimate from limited FOV RGB images. Our SMD convolutional modules can customize the network parameters to adapt to lighting conditions of a single input, and will enforce the network to implicitly learn the differentiation of various lighting environments and camera responses.

3 Method

3.1 Network Architecture Design

To realize the objective of recovering a panoramic HDR environment map from an LDR photograph, we utilize a fully convolutional encoder-decoder network architecture (Figure 2) as the general backbone of our network structure (the baseline network). We choose to ultilize ResNet followed by an adaptive average pool layer as the encoder. This component is flexible and can be any other popular feature extractor such as VGG [34] and DenseNet [18]. Since the output is represented as a spherical panorama, we use spherical convolutions proposed by Coors et al. [7] to correctly weight non-uniform polar distortions (e.g., overhead lighting) in equirectangular images as well as address boundary discontinuities on the left and right-most pixels (thus removing seam artefacts).

One of our design goals is to learn lighting-specific features hidden within the input photos. We propose to use several dynamic convolutional modules as the refinement block to model input-adaptive features and adaptively interpret the features while decoding in the spherical domain. We embed Spherical Multi-scale Dynamic (SMD) convolutional modules to the first three layers of decoder since the features are most dense around these layers, and use a Multi-scale Dynamic (MD) module with normal 2D convolutions at the end layer of encoder to further refine extracted features. These dynamic convolutional modules use the general feature maps extracted from the general convolutional layers to dynamically generate convolutional filters. These dynamic filters can extract feature maps customized for representing specific lighting variations from the input sample. In these dynamic modules, we use a multi-scale scheme that generates different size dynamic kernels from general feature maps. This captures features of different scales that are relevant to the specific input, making the entire dynamic module more flexible for both coarse and fine features. Multi-scale dynamic features are extracted in parallel and then concatenated with general features together as the complementary features.

The training process can be formulated as equation 1. I is the input LDR photograph that has limited FOV. G_h is the ground truth of the 360° HDR environment map. $f(\cdot)$ represent the network that maps the LDR photograph to the HDR environment map. Θ represents the parameters of the network that need to be learned and optimized.

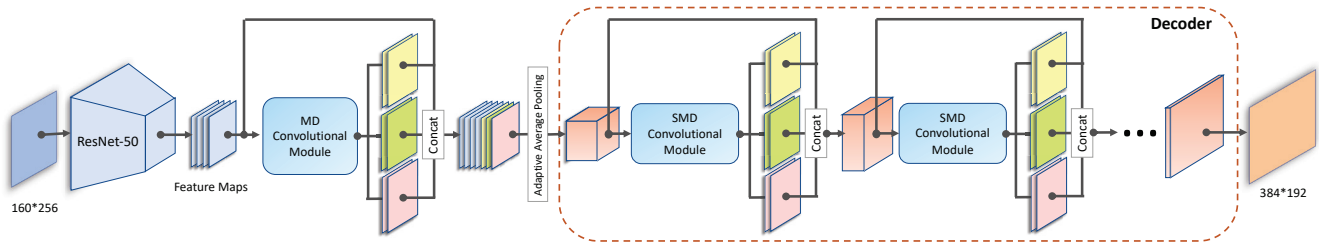


Fig. 2: The overall structure of DLNet. A Multi-scale Dynamic (MD) convolutional module extracts lighting-specific features from the feature maps generated by the encoder using a ResNet-50 model. The concatenated feature maps are then fed to the decoder after adaptive pooling, where the first 3 decoding layers uses Spherical Multi-scale Dynamic (SMD) convolutional modules to perform input-specific feature interpretation. Finally, a panoramic HDR environment map is generated by the decoder.

$\mathscr{L}$ represents the loss function, where we use RMSE. In DLNet, the parameters are split into two categories: shared parameters Θ_s from the backbone network that learns the general features shared among all data, and dynamic parameters Θ_d which learns the dynamic features specific to each individual input. We optimized them together simultaneously in the training process

$$\Theta^* = \arg\min_{\Theta_s,\Theta_d} \mathscr{L}(G_h;\ f(I;(\Theta_s;\Theta_d))) \tag{1}$$

When combining dynamic convolution modules into the backbone network, different normalizations are needed to balance between dynamic features and general features. Dynamic features are more instance-oriented and are more suitable to use with instance normalization, while the general features may benefit from batch normalization or layer normalization. We choose to use switchable normalization (SN) [25] in the entire network including the encoder. The neural network can learn the optimal combination of batch normalization, instance normalization and layer normalization in an end-to-end manner to adapt to DLNet structure. We use spherical convolution [7] for both general networks and dynamic modules to address nonuniform distortion caused by the equirectangular image.

3.2 Spherical Multi-scale Dynamic Convolutional Modules

The aim of the SMD convolutional module is to model the discrepancies between shared and input-customized lighting features dynamically conditioned on each input sample. It consists of two parts: the dynamic filter generator and dynamic convolutional layer. The general feature maps extracted from the backbone network are used as the input of the dynamic filter generator, and will be transformed to the input-customized dynamic filters. In the training process, the dynamic filter generator will learn a set of parameters to generate these filters rather than directly learn the filter parameters. In the testing process, the module generates suitable filters dynamically for each input rather than use fixed trained parameters of filters found in traditional CNNs.

The dynamic convolutional layer takes the general feature maps from the backbone network as the input and convolves them with the dynamic filters to obtain the input-specific feature maps, as shown in Figure 3. These features will be more selective to lighting conditions of each input samples and complement the general features captured by the backbone network.

The dynamic filter generator, which is presented as $g(F)$ in equation 2, has multiple convolutional layers and an adaptive max pooling layer. In order to learn the input-specific lighting variations from fine to coarse, we propose a multi-scale framework using different branches to capture different scales of lighting features conditioned on each input photograph. Here we use a kernel size of $\{1\times1, 3\times3, 5\times5\}$ as the filter sizes of three branches in each dynamic filter generator. In the dynamic convolutional layer, the feature maps F extracted from the backbone network first get channel-wise reduced to F_r using point wise convolution and then spherically convolved with the generated dynamic filters Θ_d to get the dynamic feature maps. The final feature maps $F^{'}$ that output to the next layer in the backbone network will be the

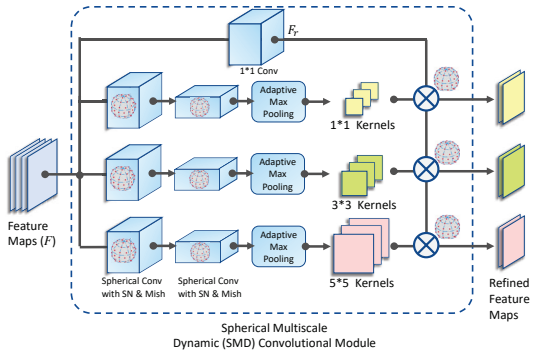


Fig. 3: Spherical Multi-scale Dynamic (SMD) convolutional module. This performs spherical convolutions on the general feature maps to generate filters of different sizes. Then the filters are spherically convolved with the processed general features maps to produce adaptive features at different scales. The Multi-scale Dynamic (MD) module has the same structure with normal convolution operations.

combination of dynamic feature maps and original feature maps. Note that the Multi-scale Dynamic (MD) module added to the end of encoder is using normal convolutional operations since it is dealing with 2D feature maps.

$$\Theta_d = \{g_{k_{1\times1}}(F); g_{k_{3\times3}}(F); g_{k_{5\times5}}(F)\}, \tag{2}$$

$$F^{'} = \{F; F_r \otimes \Theta_d\}, \tag{3}$$

Specifically, DLNet has the input size of (160,256) and output size of (384,192). The setups of each layer are listed in Table 1. The latent space has 2048 channels with feature maps the size of (3,6). There are 6 decoder layers in all. For each decoder layer, bilinear upsampling is followed by a convolution layer to scale the output size up to 2 times of the input. In SMD/MD convolutional module, each dynamic filter generator network includes 2 convolutional layers before an adaptive max pooling layer, of which each convolutional layer has a kernel size of 3, stride of 2 and $\frac{1}{4}$ input channels as the output number of channels. Each dynamic convolutional layer has a kernel size of 3 and stride of 1. Switchable normalization and mish activation functions [27] are used on all encoder and decoder layers including the SMD modules.

3.3 Data Augmentation

We use the Laval HDR indoor dataset as our training dataset [11]. This is a dataset of HDR environment maps of which we can crop out photographs to obtain input/output pairs for training. Based on prior work [4, 11], we crop out 8 images with different view directions (uniformly distributed along the horizon line) for each HDR environment map, which are then tone mapped to LDR as the input. The zenith

Layer	Kernel Size	Stride/Scale	Channel	Output (h,w)
Resnet50	default	default	2048	(8,5)
Adaptive Average Pool	-	-	2048	(3,6)
MD_0	$(1\times1,3\times3,5\times5)$	(2,2,1)	$2048+3*512$	(6,12)
decoder_1	(3×3)	(2)	1024	(6,12)
SMD_1	$(1\times1,3\times3,5\times5)$	(2,2,1)	$1024+3*256$	(6,12)
decoder_2	(3×3)	(2)	512	(12,24)
SMD_2	$(1\times1,3\times3,5\times5)$	(2,2,1)	$512+3*128$	(12,24)
decoder_3	(3×3)	(2)	256	(24,48)
SMD_3	$(1\times1,3\times3,5\times5)$	(2,2,1)	$256+3*64$	(24,48)
decoder_4	(3×3)	(2)	128	(48,96)
decoder_5	(3×3)	(2)	64	(96,192)
decoder_6	(3×3)	(2)	3	(192,384)

Table 1: DLNet network details, including the modules, setups for input and output dimensions, kernel sizes, strides and channel numbers. For the SMD/MD convolutional modules (MD_0, SMD_1, SMD_2, SMD_3), we use a stride of 2 in the first 2 convolution layers and a stride of 1 for spherical convolution layer.

Variations	Values
Color temperature (K)	[2850, 3800, 5500, 6500, 7500]
Exposure	$[0.25, 0.5, 0.75, 1.42, 1.83, 2.25]$
FOV (degrees)	$[60, 70, 80, 90, 100, 110, 120, 130]$

Table 2: Data augmentation variations in color temperature, exposure, and FOV.

angle of the crop is chosen randomly between 90° and 135° (looking downward, typical in AR applications where objects are on the floor). We also choose a point if it is lying on a valid flat plane (via normal detection). Similar to Gardner et al. [11], each environment map was warped and re-centerd such that the crop is in the center so that the lighting better represented the illumination that would be incident at the position where the object would be composited as opposed to the 360° camera's position that can be arbitrarily far away. This operation ensures that the CNN learns the correlation between illumination cues within the crop and its corresponding environment map at the correct point of composition. Otherwise, the lighting mismatch between the input crop and target environment map would make the training process challenging. To make the dynamic range of environment maps controllable for network training, we normalized the data range in log space using $0.2*(G_h)^{\frac{1}{2.2}}$ similar to [42]. The estimation is re-scaled to the full range using the inverse function $\frac{1}{0.2}(G_h{}^*)^{2.2}$. There are 16464 and 688 samples for the training and testing dataset prior to data augmentation.

To ensure our model is robust to lighting variation and sufficiently practical to real-world use cases, we create data augmentations by considering the wide variation of lighting conditions imposed by the environment, as as well how the camera sensor responds to the environment. We choose our augmentations based on available consumer level devices, literature, and experimentally. In our experiments, we found three properties that greatly impact lighting and light estimation: color temperature, exposure, and FOV. The augmentation values for each are summarized in Table 2.

Color temperature is a measure expressed in Kelvins (K) for blackbody radiation where the wavelengths correspond to "warm" to "cool" color tones. This is evident in HDR environment maps as well [5]. Furthermore, color casts can be caused by the camera's automatic white balance. We use prior work [1, 2] to generate white balance augmentations that are similar to color casts caused by cameras. This augmenter produces images with a target Kelvin value of $[2850K, 3800K, 5500K, 6500K, 7500K]$. We apply this to the cropped image. This assumes that if the image has a color cast, then any virtual objects composited into it will be illuminated by an estimated light source that produces a similar color cast on the virtual object. We apply color temperature changes to the photo's corresponding environment map as well. The intensity of lighting depends on the lumens of a light bulb as well as the aperture of the camera. Based on prior work [6], we experimentally chose a lower bound of 0.25 and 2.25 scalar intensity value which corresponds a well under and overexposed image. We choose equally spaced steps between $[0.25, 1]$ and $[1, 2.25]$, resulting in the following exposure values: $[0.25, 0.5, 0.75, 1.0, 1.42, 1.83, 2.25]$.

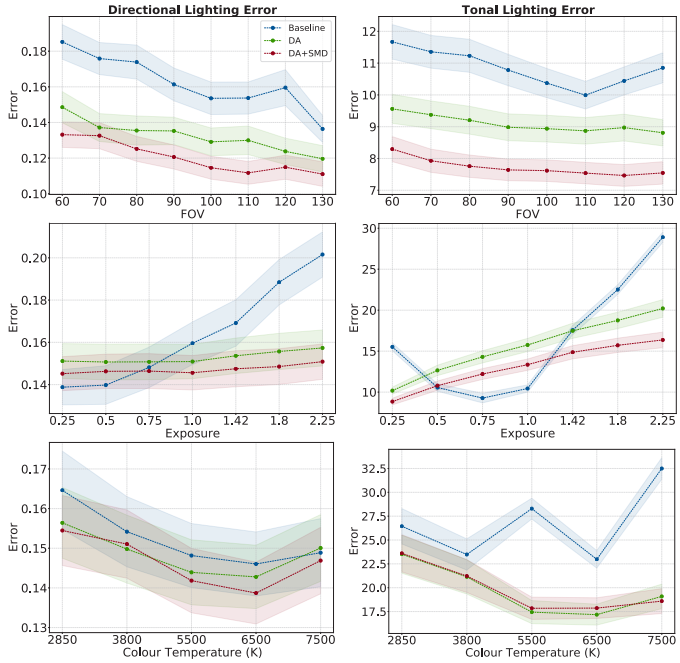


Fig. 4: Ablation study with respect to FOV (top), exposure (middle), and color temperature (bottom) variation (95% CI). Blue is the baseline, green is the DA model, and red is the DA+SMD model.

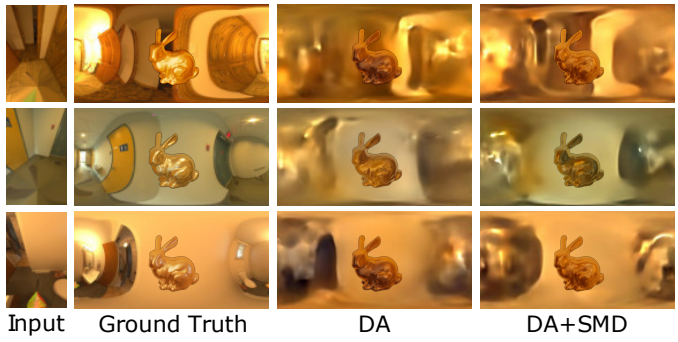


Fig. 5: Examples from the ablation study. We show the environment maps along with a virtual bunny illuminated by its respective environment map in the center. We can observe how SMD improves in defining details on both environment map and rendered object in terms of light source intensity, direction, and overall structure.

Finally, for FOV, after surveying cameras on popular mobile devices, we found an FOV range between 70° to 120°. We therefore produce a dataset to encapsulate this range with the following FOV values: $[60, 70, 80, 90, 100, 110, 120, 130]$. While our network is trained on pre-defined augmentation intervals, in our experiments we found that the network did not have issues adapting to in-between (interpolated) values.

3.4 Training Schedules

We use Root Mean Squared Error (RMSE) loss function and Adam optimizer to train the model. The learning rate is 0.005 and the batch size is 16. We train 200 epochs on any augmented dataset and 500 epochs on the non-augmented dataset. We observe that the models are well converged within half of the number of epochs. Using an Nvidia Quadro RTX 6000 GPU cards with 24GB of memory, it takes around 0.5 hours for the non augmented dataset for each epoch.

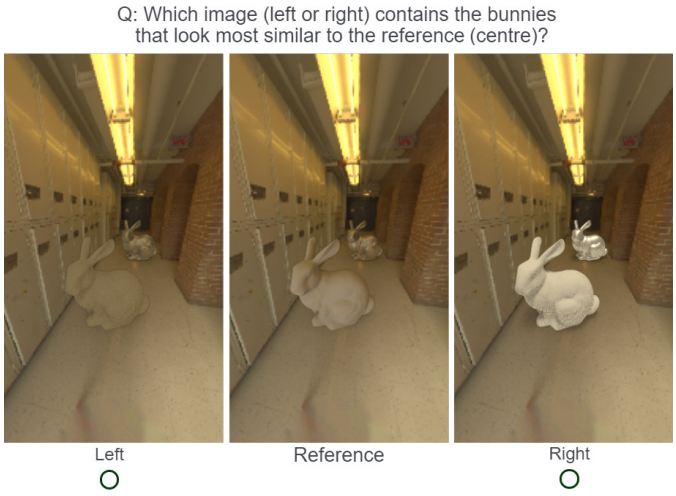


Fig. 6: Sample question from the user study. Each image contains two rendered bunnies. The left and right images are illuminated with the estimated lighting, and the center image contains the ground truth lighting. Users choose between the left and right image based on the question above the images.

4 Experiments

Dataset. We used the 688 photographs from our test. These photographs are augmented across FOV, exposure, and color temperature, bringing the total to 27,273 test photographs. Each photograph has a corresponding ground truth HDR environment map which is used to measure the error from the estimates. For the ablation study, the default FOV is 120°. The default exposure and color temperature are the original versions from the dataset (no augmentation).

Metrics. We use metrics in both the ablation study and comparison with prior work. Two metrics are used to measure important properties of environmental lighting: *"Directional Lighting Error"* and *"Tonal Lighting Error"*. The directional metric [4] measures the angular error between the nearest pairs of lights (via light detection [33]) between the ground truth and the estimate. The value is normalized between 0 and 1. Intuitively, a value of 1 corresponds to a maximum angular error of 180°. The tonal lighting measure computes the mean of the environment map for each color channel (linear RGB), and converts the RGB value into LAB (D65) color space [28]. Note that the environment maps in our database are stored in relative units. We calibrate the data by applying an auto-exposure to each environment map such that it's mean is equal to 0.2. The error is then computed by taking the Euclidean distance between the mean LAB colors of the ground truth and the estimate. The mean of the environment map is meaningful in a rendering sense, as it is equivalent to taking the mean of a diffuse map. Note that due to human perception being less sensitive to darker intensities, the results will show lower error for lower exposures. This effect could be removed through normalizing the data, but we omit this step to adhere to human perception.

In our evaluation, we use these two metrics on the dataset to obtain the mean directional lighting and tonal lighting error. We also calculate the 95% confidence intervals (CI) around the mean, which is shown with different color blocks around each line in Figure 4 and Figure 8.

Prior Methods. In our quantitative and qualitative results, we compare our method with the state-of-the-art methods in light estimation from a single RGB indoor photograph. For this, we chose the seminal work by Gardner et al. [11], and two other recent approaches by LeGendre et al. [23] and Chalmers et al. [4] (each denoted as [Gardner17], [LeGendre19] and [Chalmers20] respectively). The approach by Gardner et a. [10] (denoted as [Gardner19]) also has been compared with in the quantitative comparison.

Runtime. All the tests have been done on one GeForce RTX 3090 card with a runtime of 0.092s (average across 100 samples) with an input image size of (160,256) and output size of (384,192).

Table 3: Ablation study for the different number of kernels in SMD/MD module (on the FoV variation dataset).

Mean Error	DA	DA + SMD (1×1)	DA + SMD (1×1, 3×3)	DA + SMD (1×1, 3×3, 5×5)
Directional Lighting Error	0.132	0.124	0.123	**0.120**
Tonal Lighting Error	9.10	9.02	8.89	**7.72**

Table 4: Ablation study for the different number of SMD/MD layers (on the FoV variation dataset).

Mean Error	DA	DA + SMD (1 layer)	DA + SMD (2 layers)	DA + SMD (3 layers)	DA + SMD (4 layers)
Directional Lighting Error	0.132	0.130	0.130	0.126	**0.120**
Tonal Lighting Error	9.10	9.08	8.48	8.21	**7.72**

Table 5: The average directional lighting error and tonal lighting error with respect to FOV, exposure and color temperature variations for both ablation studies and the state of the art (95% CI).

Mean Error	Directional Lighting Error			Tonal Lighting Error		
	FOV	Exposure	Temperature	FOV	Exposure	Temperature
Baseline	0.159±0.009	0.164±0.009	0.156±0.009	10.72±0.48	17.81±0.57	26.74±1.33
DA	0.130±0.008	0.153±0.008	0.149±0.008	9.02±0.42	15.54±0.79	**19.67±1.48**
DA+SMD	**0.119±0.007**	**0.147±0.008**	**0.147±0.008**	**7.64±0.35**	**13.07±0.71**	19.83±1.44
[Chalmers20]	0.144±0.007	0.161±0.008	0.205±0.009	11.50±0.49	19.96±0.54	29.54±1.3
[LeGendre19]	0.444±0.017	0.416±0.015	0.468±0.017	22.83±1.05	21.57±1.01	31.63±1.45
[Gardner17]	0.345±0.016	0.327±0.016	0.314±0.015	25.01±1.20	30.95±1.19	30.47±1.39
[Gardner19]	0.166±0.009	0.161±0.008	0.184±0.009	43.34±0.84	42.38±0.83	47.23±1.24

4.1 Ablation Study

We have addressed photograph variation by using DLNet which uses SMD modules within the network architecture as well as data augmentation as part of the dataset. To understand the contribution of data augmentation and the SMD modules, we demonstrate how our network performs when using data augmentation alone (denoted as DA) as well as using DA and SMD together (denoted as DA+SMD) against the baseline network without either of them. The DA network concatenates all the components listed in Table 1 except the SMD/MD convolutional modules, and uses the ResNet-pretrained model on the ImageNet dataset [9] for encoder initialization, with spherical convolutions in all decoder layers and switchable normalization in entire network as described in Section 3.1. The DA+SMD network concatenates all the components listed in Table 1 while keeping other components the same as DA. The baseline network architecture is the same as DA. The results are shown in Figure 4.

We observe that our result improves upon the baseline in almost all cases. Further, the baseline shows inconsistencies across different variations while both DA and SMD improves it by with more stability. The baseline model is often biased toward specific points, depending on the original dataset distribution. For example, the baseline is biased toward the original exposure when measured with tonal lighting error metrics shown in Figure 4 (column 2, row 2). Since the baseline model was trained with only the original auto-exposed data with no variation, it over fits to the data with similar exposure values. We also see in Figure 4 (column 1, row 2), that as exposure increases, the baseline's light direction error increases too. The baseline will overestimate additional light sources as exposure increases, whereas it should instead fix the estimated number of lights and vary the overall exposure of the estimated environment map. The DA and SMD helps produce more generalized and robust results for different variations. DA improves the baseline performance across all variations. Based on DA, the SMD module shows further improvements with respect to varying FOV (Figure 4, row 1) and exposure (Figure 4, row 2).

Interestingly, we observe that DA by itself handles shifts in color temperature (row 3). Due to this, we believe that DA aids the network understanding changes in color, but the SMD module aids the network in defining details (structural details, direction and intensity of light sources). We illustrate this with an example in Figure 5 which shows the environment maps and a corresponding virtual bunny illuminated by the environment map. From the environment map estimation, we can observe the general ambient tones are similar between DA and

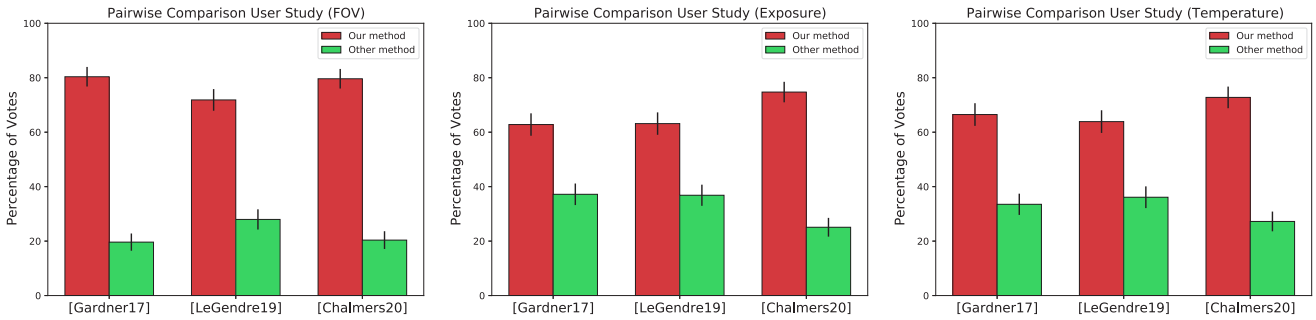


Fig. 7: Pairwise user study between our method and the state of the art with respect to FOV (left), exposure (middle), and color temperature (right). Higher is better. Red is our method, and green is the state-of-the-art *"other method"*. The labels below each pair of bars indicates what the *"other method"* is referring to.

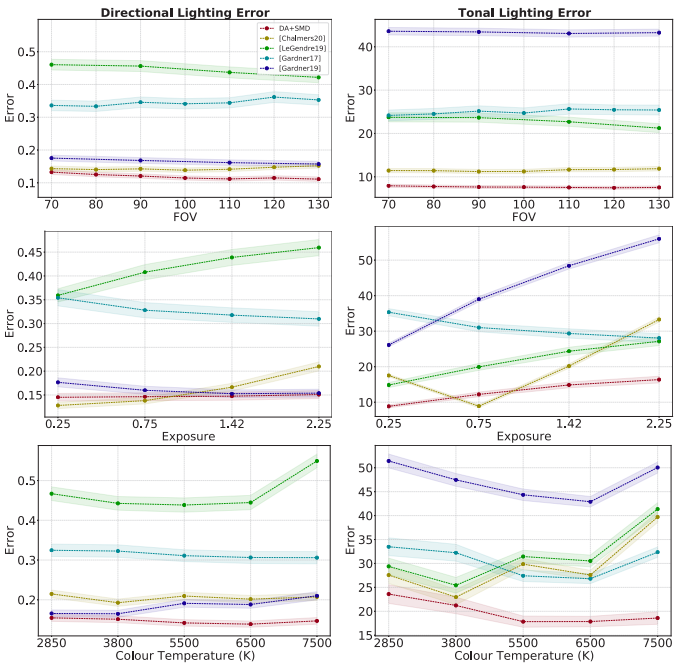


Fig. 8: Quantitative state-of-the-art comparisons with respect to FOV (top), exposure (middle), and color temperature (bottom) variation (95% CI). Blue is [Gardner19], cyan is [Gardner17], green is [LeGendre19], yellow is [Chalmers20], and red is our DA+SMD model.

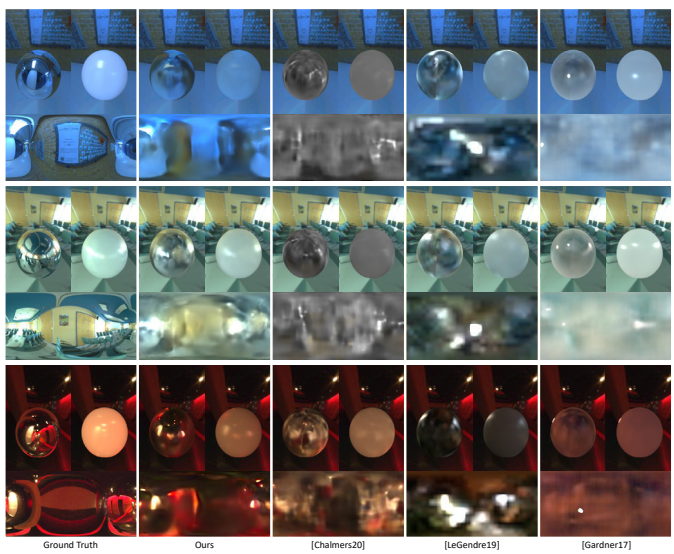


Fig. 9: Qualitative state-of-the-art comparisons with three different input photographs (each row). For each case, we composite a mirror sphere and glossy sphere. Below each sphere rendering are the environment map used to illuminate the spheres.

DA+SMD, but the SMD module is able to bring out the light sources with more detail and in the correct direction. The illuminated bunnies also shows this effect, for example, the highlights on the chest region of the DA+SMD bunny in row 1 and the body and cheek region of DA+SMD bunny in row 3. The light sources from the DA bunnies are also duller in general than DA+SMD in all three rows. The overall structure in the environment maps are also improved by SMD (well-defined edges, improved sense of depth).

We also did an ablation study on the different number of kernels in SMD/MD modules, as well as different SMD/MD layers in the network structure on FOV variations. For the experiments of kernel number in SMD/MD modules, we include all SMD/MD layers listed in Table 1 from MD_0 to SMD_3 in the network and append kernels from 1×1, 3×3, 5×5 one at a time. For the experiments of SMD/MD layers, we keep all 3 kernels in each SMD/MD layer and append SMD/MD layers one by one. The mean error results for all FOV variations (listed in Table 2) are shown in Table 3 and Table 4. We can observe that more kernels and SMD layers improves both directional and tonal lighting error.

4.2 Comparison

A user study and metrics are used to quantitatively evaluate our method.

Comparison with a user study. Similar to prior work [11, 12], we conducted a pairwise comparison user study to evaluate how well the light estimate matches the ground truth. Users were presented pairs of images containing virtual objects (two Stanford bunnies with a diffuse and metallic material respectively composited into the photo with a shadow catcher) illuminated by lighting produced our method and one of the state-of-the-art methods. Using the ground truth as a reference, participants were asked to select which image had virtual objects that is most visually similar to the reference (i.e. 3 images are shown for each case, users make judgments only between left and right images, the middle image (ground truth) is for their reference only). We selected 10 images from the test dataset, and generated 3 variations for each of the 3 different properties (FOV, exposure, colour temperature) for a total of 90 images. For each image, participants answered 3 pairwise comparisons (our method against 3 state-of-the-art methods). We had a total of 19 participants, aged between 18 and 64, 26.3% female and 73.7% male. A sample question from the study is shown in Figure 6. A summary of statistical results including user responses, CI and p-values (denoted as p) analyzed using a binomial test is summarized in Table 6 and Figure 7 (with 95% CI error bars). From the results, we can see that the majority of participants preferred our method over prior work

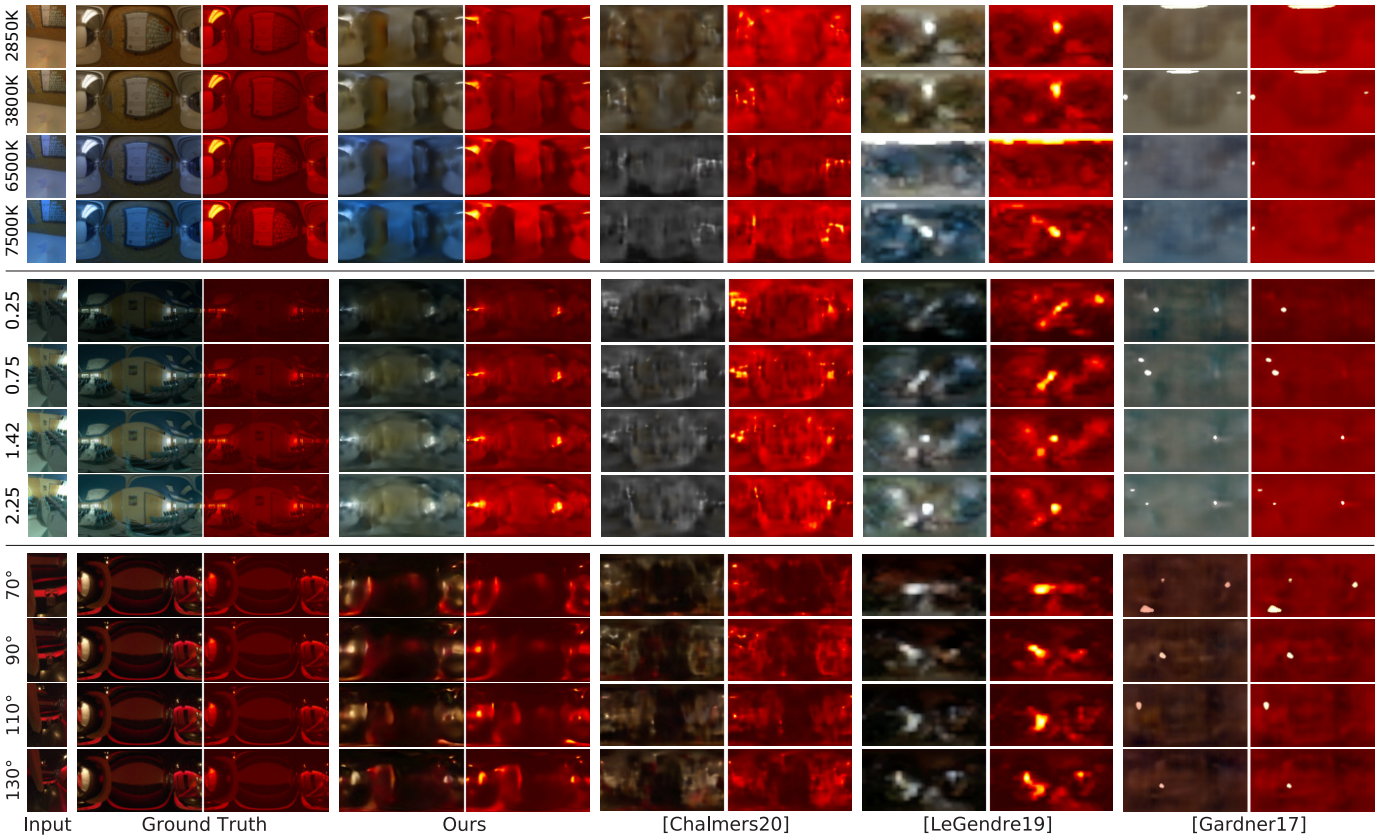


Fig. 10: Qualitative comparison. The left column is the input photograph, and to the right is the ground truth, our method, and the state-of-the-art's lighting. Beside each environmental map is its corresponding heatmap. Vertically we adjust the variations for each category: color temperature (top), exposure (center), and field-of-view (bottom).

across all properties with statistical significance ($p<0.001$). We observe this result across all variations (FOV, exposure, and colour temperature). Between the variations, our method was particularly more preferred in FOV (77%), followed by both exposure (67%) and temperature (67%).

We also conducted a pairwise comparison, between our method and the ground truth (the two images are side-by-side, users are asked to choose which image has the rendered bunnies that look more realistic - using the background photograph for the reference illumination). The same group of participants joined in this user study. The results from our study showed that users chose our method as more realistic $36.9\% \pm 4.0$, $31.6\% \pm 3.9$, $38.7\% \pm 4.0$ of the time for FOV, exposure and colour temperature respectively (with 95% CI), while the ground truth was chosen $63.1\% \pm 4.2$, $68.4\% \pm 4.2$, $61.3\% \pm 4.2$ of the time (with 95% CI). A confusion rate of 50% is ideal as it would indicate that the ground truth and our lighting estimate would be difficult to distinguish between. Our results indicate that DLNet is comparable with the ground truth particularly in colour temperature and FOV, but still yielded good results for exposure as well. Users will often confuse our estimate and the ground truth with one another. However, when the cropped images have ambiguous lighting cues (not enough scene information), DLNet will not be able to predict the correct light sources direction which will make rendered images not as good as the ground truth. Our result is also still acceptable based on prior work who conducted a similar study [11, 23].

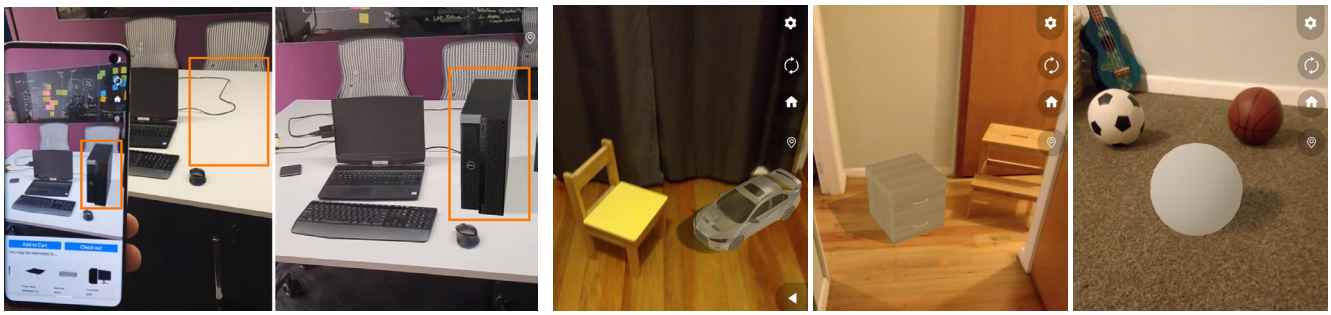
Comparison using metrics. The results using the metrics are shown in Figure 8. Note that we obtained only four FOV variations for [LeGendre] (generously provided by LeGendre et al.). We observe that our method outperforms the state of the art in FOV (row 1) and color temperature (row 3) for both metrics, and in most cases for exposure (row 2) for both metrics. Interestingly, we can also observe that in terms of directional lighting (column 1), our method is able to retain a similar error value across all variations. This indicates that our light estimates are structurally consistent when the textural content within the photograph is the same.

We listed the average directional lighting error and tonal lighting error for both the ablation study and prior work comparisons in Table 5. Compared with prior work, our baseline already shows competitive results, although not good as [Chalmers20] in terms of directional lighting error. The baseline benefits largely from the pre-trained ResNet encoder, and the spherical convolutions in all the decoder layers helps address the distortion in equirectangular environment maps. It also benefits from the switchable normalization and mish activation, all of which have been shown to be able to boost the modeling capability. When using augmented data to train the baseline network, we are able to get the DA model superior than the other works on both metrics, especially on color temperature variations. Overall, DA+SMD achieves the best performance in most cases due to the SMD's ability to adapt to the wide range of input variations.

We also use the average RMSE to measure the reconstruction error of the environment map. The mean values of different FOV, exposure and color temperature are reported in Table 7. It shows that our method has lower average RMSE than [LeGendre19] and [Gardner17], and shows slight improvements over [Chalmers20] and [Gardner19].

4.3 Qualitative Results

Figure 1 shows an example of a single scene with variations of color temperature, exposure, and FOV, where our method closely matches the ground truth. The structure of the estimate, including light direction, is retained across varying color and exposure. The estimate is also robust with increasing FOV, increasing with accuracy. We show qualitative comparisons using rendered spheres in Figure 9, and environment maps in Figure 10. We also visualize alongside each environment map its corresponding heatmap to better illustrate where the high-intensity areas of light are, and how bright they are relative to the rest of the HDR



(a) AR Application (left: over-the-shoulder view, right: fullscreen view).

(b) Additional examples (fullscreen view).

Fig. 11: Our light estimation model integrated into an AR/MR framework. (a) The orange rectangle indicates where the virtual object (desktop computer) is placed. Left: Over-the-shoulder view of a user placing a virtual desktop computer into the scene. Right: Fullscreen view of the mobile display. (b) Additional examples (from the left: virtual car, virtual drawers, virtual diffuse sphere).

Table 6: Statistical results on FOV, exposure and color temperature for pairwise comparisons between our method vs. [Gardner17], our method vs. [LeGendre19], and our method vs. [Chalmers20] with ground truth as a reference, as well as a direct pairwise comparisons between our method and ground truth (Our method vs. GT). The p-values in bold indicates statistical significance. The mean response score is provided, along with 95% confidence intervals (CI).

		Our method vs. [Gardner17]	Our method vs. [LeGendre19]	Our method vs. [Chalmers20]	Our method vs. GT
FOV	Response	(**80.4**% ± 3.6) vs. (19.6% ± 3.2)	(**71.9**% ± 4.0) vs. (28.0% ± 3.7)	(**79.6**% ± 3.6) vs. (20.4%±3.3)	(36.9%±4.0) vs (**63.1**% ± 4.2)
	p-value	**p<0.001** (p = 4.37e-48)	**p<0.001** (p=4.72e-25)	**p<0.001** (p=1.12e-45)	**p<0.001** (p=1.95e-10)
Exposure	Response	(**62.8**% ± 4.1) vs. (37.2% ± 4.0)	(**63.2**% ± 4.1) vs. (36.8% ± 3.9)	(**74.7**% ± 3.8) vs. (25.1%±3.4)	(31.6% ± 3.9) vs. (**68.4**%±4.2)
	p-value	**p<0.001** (p = 1.02e-09)	**p<0.001** (p=3.46e-10)	**p<0.001** (p=1.24e-33)	**p<0.001** (p=2.24e-18)
Temperature	Response	(**66.5**% ± 4.2) vs. (33.6% ± 3.9)	(**63.9**% ± 4.2) vs. (36.1% ± 4.0)	(**72.8**% ± 4.0) vs. (27.2%±3.6)	(38.7% ± 4.0) vs. (**61.3**%±4.2)
	p-value	**p<0.001** (p = 1.51e-14)	**p<0.001** (p=1.1e-10)	**p<0.001** (p=6.8e-27)	**p<0.001** (p=3.43e-09)

Table 7: Average RMSE between the predicted environment map (ours and the state of the art) against the ground truth with respect to FOV, exposure, and color temperature variations (95% CI).

Average RMSE	Our method	[Chalmers20]	[LeGendre19]	[Gardner17]	[Gardner19]
FOV	**2.18±0.23**	2.21±0.23	2.30±0.24	3.92±0.25	2.29±0.23
Exposure	**2.83±0.35**	2.84±0.35	3.47±0.34	4.93±0.36	2.87±0.35
Color Temperature	**6.74±1.44**	6.85±1.44	7.88±1.43	8.34±1.42	6.87±1.44

environment map (as opposed to regular RGB which clip the lights to white). The data was nomalized into range [0,1] in the visualization. We also show our method working in an interactive AR/MR framework (Figure 11), where our lighting estimate provides high visual quality results for images in the wild. In Figure 11-(a), the virtual desktop computer matches the shadow direction of the real laptop. Additional examples are shown in Figure 11-(b). The virtual car (left), drawer (center), and diffuse sphere (right) match the tone and shadow direction of the real stool and balls respectively. Note that the virtual car has a glossy material that visibly reflects the real black curtain and ceiling light (half the roof reflects the black curtain, the other half contains a specular highlight from the ceiling light). We show additional examples of our light estimation working well using the interactive AR/MR framework with wide variation (FOV, exposure, colour) in Figure 12. We vary the camera exposure and color by adjusting the exposure multiplier and color filters respectively. The light estimation adapts to the changes in color/exposure while still maintaining the lighting structure (e.g., light direction). For FOV, we used three different phones with different FOV (OnePlus 5, Samsung Galaxy S10, and Samsung Galaxy S21, with 67°, 77°, and 83° FOV respectively). Note that the wider FOV images (row 1, column 2, 3 and 5, 6) has improved light estimation quality compared to the lower FOV (row 1, column 1 and 4) due to our model taking advantage of the additional information provided by the wider FOV. For different variations of each scene, we captured images standing at similar positions.

As indicated by the metrics, we can qualitatively observe that our light estimate is structurally consistent despite changes in temperature and exposure. From Figure 10-top, the light direction, general structure, and tone is much more similar to the ground truth than prior work. As the tones move from warm to cool, our light direction and structure remain intact, while the prior work's primary lighting directions shift. We can also see similar behavior with exposure variations (Figure 10-center), where our result better matches the ground truth prediction in quality, and better retains consistency in light direction and structure. The FOV is different to the color temperature and exposure variations, as the actual scene content changes. As such, we expect the light estimates to change structurally. Ideally, the light estimate should improve with wider FOV since there is more potential for lighting cues to be observed within the photograph. We observe this behavior in Figure 10-bottom, where our network provides a good estimate gradually improves the quality of the result as the FOV increases. This is as opposed to the prior work, where we don't see their result improve as the FOV increases.

5 Discussion and Limitations

The benefit of our model is that we are able to account for a wide variation of input photographs in terms of FOV, exposure and color temperature. While this benefits the overall light estimate's quality, we can also use these properties for additional contributions. Robust lighting estimation from varying exposure is able to support environmental fluctuations in intensity as well as the aperture of the device. Therefore, out-of-view lighting changes (e.g., closing a blind, cloud movement in overcast conditions) can be estimated with temporal coherency. Our supplemental video showcases a few examples of our model's capabilities.

While DLNet retains the lighting direction well across variations, small light sources still can observed shifting or fading. This may be because the lighting cues are ambiguous or subtle for small light sources. This could be improved with an extra refinement step. Irregular-shaped light sources are also challenging to estimate. Our estimation for the specular component is accurate compared with the previous work but still blurry in comparison with the ground truth, which may improve with more training examples. Data augmentations with variations in light shapes along with DLNet may address this issue. Further, in our user study we conducted the binomial test, which is based on the assumption that the samples are a fair representation of the population. Our demographic's female rate is only 26.3%, whereas it should ide-

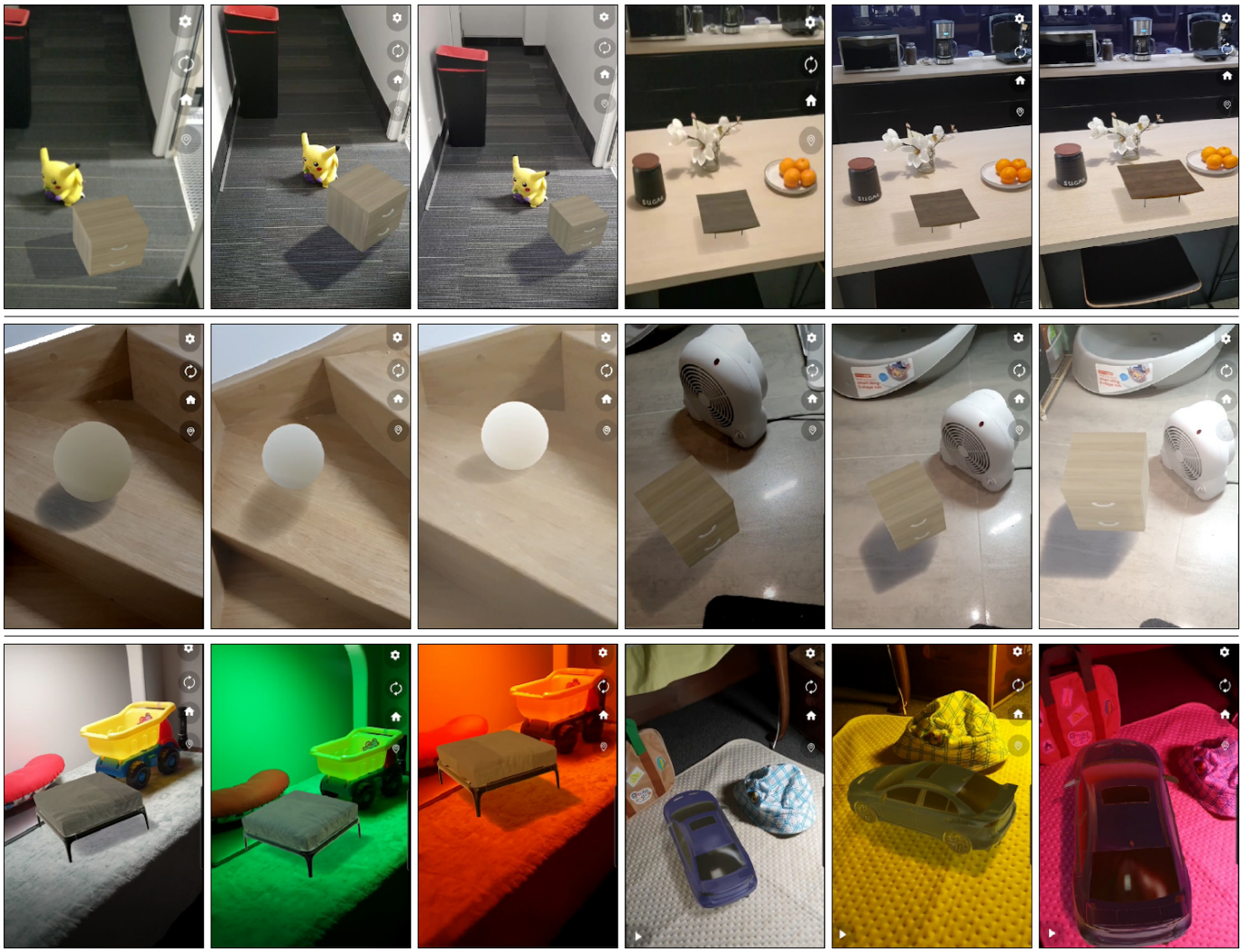


Fig. 12: AR examples in the wild with respect to FOV (row 1), exposure (row 2) and color variations (row 3). For each case, we show 2 scenes with 3 variations. Column 1-3 for one scene and column 4-6 for the other scene. For FOV and exposure cases, we show variations in a increasing order (short, medium, long, and 67 °, 77 °, and 83 °respectively). For color variations, we show the camera's default colour temperature along with two different colour variations (green, red, yellow and pink).

ally be closer to 50%. Most of the participants we had are from the engineering sector, which is comprised mostly of males. Future work should ensure the sampling better represents the population.

6 Conclusion and Future Work

We have presented DLNet and a data augmentation strategy that is able to accommodate the wide variation of photographs caused by variations of lighting in the environment (light intensities, color temperature, etc.) and variations of the sensor (FOV, exposure, and color temperature). We achieve this by using SMD modules that are able to adaptively generates the convolution filters of the network specific for each new input sample. We also generate data augmentations into the training set, encompassing the expected variations of lighting and sensors from the wild. We demonstrate the advantages of our model through an ablation study and comparison with state-of-the-art light estimation techniques. We also consider various use case scenarios that can benefit from our model. Future work can consider integrating the camera response function as part of the input. Our method could be applied to mobile AR/MR applications to improve lighting and composition. Special attention can also be given to specifically target robustness against temporal consistency.

7 Acknowledgement

This project was supported by the Smart Ideas project funded by MBIE and the Entrepreneurial University Programme funded by TEC in New Zealand.

References

[1] M. Afifi and M. S. Brown. What else can fool deep learning? Addressing color constancy errors on deep neural network performance. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 243–252, 2019.

[2] M. Afifi, B. Price, S. Cohen, and M. S. Brown. When color constancy goes wrong: Correcting improperly white-balanced images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1535–1544, 2019.

[3] H. Barrow, J. Tenenbaum, A. Hanson, and E. Riseman. Recovering intrinsic scene characteristics. *Computer Vision Systems*, 2(3-26):2, 1978.

[4] A. Chalmers, J. Zhao, D. Medeiros, and T. Rhee. Reconstructing reflection maps using a stacked-CNN for mixed reality rendering. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[5] A. Chalmers, T. Zickler, and T. Rhee. Illumination Space: A feature space for radiance maps. *Proceedings of the Pacific Graphics*, 2020.

[6] D. Cheng, J. Shi, Y. Chen, X. Deng, and X. Zhang. Learning scene illumination by pairwise photos from rear and front mobile cameras. In

Computer Graphics Forum, vol. 37, pp. 213–221. Wiley Online Library, 2018.

[7] B. Coors, A. Paul Condurache, and A. Geiger. SphereNet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 518–533, 2018.

[8] 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 ACM SIGGRAPH*, 1998.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

[10] M.-A. Gardner, Y. Hold-Geoffroy, K. Sunkavalli, C. Gagné, and J.-F. Lalonde. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7175–7183, 2019.

[11] 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)*, 9(4), 2017.

[12] M. Garon, K. Sunkavalli, S. Hadap, N. Carr, and J.-F. Lalonde. Fast spatially-varying indoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6908–6917, 2019.

[13] S. Georgoulis, K. Rematas, 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*, 40(8):1932–1947, 2017.

[14] C. Han, S. Shan, M. Kan, S. Wu, and X. Chen. Face recognition with contrastive convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 118–134, 2018.

[15] J. He, Z. Deng, and Y. Qiao. Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3562–3572, 2019.

[16] Y. Hold-Geoffroy, A. Athawale, and J.-F. Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6927–6935, 2019.

[17] Y. Hold-Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J.-F. Lalonde. Deep outdoor illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7312–7321, 2017.

[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

[19] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool. Dynamic filter networks. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NIPS)*, pp. 667–675, 2016.

[20] Y. Jo, S. Wug Oh, J. Kang, and S. Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3224–3232, 2018.

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

[22] J.-F. Lalonde and I. Matthews. Lighting estimation in outdoor image collections. In *2014 2nd International Conference on 3D Vision*, vol. 1, pp. 131–138. IEEE, 2014.

[23] C. LeGendre, W.-C. Ma, G. Fyffe, J. Flynn, L. Charbonnel, J. Busch, and P. Debevec. DeepLight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5918–5928, 2019.

[24] Z. Li, M. Shafiei, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2475–2484, 2020.

[25] P. Luo, J. Ren, Z. Peng, R. Zhang, and J. Li. Differentiable learning-to-normalize via switchable normalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[26] M. Maximov, L. Leal-Taixé, M. Fritz, and T. Ritschel. Deep appearance maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8729–8738, 2019.

[27] D. Misra. Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681*, 2019.

[28] K. Moreland. Diverging color maps for scientific visualization (expanded). *Proceedings in ISVC*, 9:1–20.

[29] T. Nestmeyer, J.-F. Lalonde, I. Matthews, and A. Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5124–5133, 2020.

[30] J. Park, H. Park, S.-E. Yoon, and W. Woo. Physically-inspired deep light estimation from a homogeneous-material object for mixed reality lighting. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):2002–2011, 2020.

[31] 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*, pp. 4508–4516, 2016.

[32] P. Ren, Y. Dong, S. Lin, X. Tong, and B. Guo. Image based relighting using neural networks. *ACM Transactions on Graphics (ToG)*, 34(4):1–12, 2015.

[33] T. Rhee, L. Petikam, B. Allen, and A. Chalmers. Mr360: Mixed reality rendering for 360 panoramic videos. *IEEE transactions on visualization and computer graphics*, 23(4):1379–1388, 2017.

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[35] S. Song and T. Funkhouser. Neural Illumination: Lighting prediction for indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6918–6926, 2019.

[36] P. P. Srinivasan, B. Mildenhall, M. Tancik, J. T. Barron, R. Tucker, and N. Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8080–8089, 2020.

[37] T. Sun, J. T. Barron, Y.-T. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P. E. Debevec, and R. Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 38(4):79–1, 2019.

[38] J. Wu, D. Li, Y. Yang, C. Bajaj, and X. Ji. Dynamic filtering with large sampling field for convnets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 185–200, 2018.

[39] Z. Xu, K. Sunkavalli, S. Hadap, and R. Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018.

[40] F. Zhan, Y. Yu, R. Wu, C. Zhang, S. Lu, L. Shao, F. Ma, and X. Xie. GMLight: Lighting estimation via geometric distribution approximation. *arXiv preprint arXiv:2102.10244*, 2021.

[41] F. Zhan, C. Zhang, Y. Yu, Y. Chang, S. Lu, F. Ma, and X. Xie. Emlight: Lighting estimation via spherical distribution approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[42] J. Zhang and J.-F. Lalonde. Learning high dynamic range from outdoor panoramas. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4519–4528, 2017.

[43] J. Zhang, K. Sunkavalli, Y. Hold-Geoffroy, S. Hadap, J. Eisenman, and J.-F. Lalonde. All-weather deep outdoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10158–10166, 2019.

[44] H. Zhou, S. Hadap, K. Sunkavalli, and D. W. Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7194–7202, 2019.