Pacific Graphics 2018
H. Fu, A. Ghosh, and J. Kopf
(Guest Editors)

Volume 37 (2018), Number 7

Learning Scene Illumination by Pairwise Photos from Rear and Front Mobile Cameras

Dachuan Cheng[1,4], Jian Shi[2], Yanyun Chen[1], Xiaoming Deng[3] and Xiaopeng Zhang[2]

[1]State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences
[2]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[3] Beijing Key Laboratory of Human Computer Interactions, Institute of Software, Chinese Academy of Sciences
[4] University of Chinese Academy of Sciences

Abstract

Illumination estimation is an essential problem in computer vision, graphics and augmented reality. In this paper, we propose a learning based method to recover low-frequency scene illumination represented as spherical harmonic (SH) functions by pairwise photos from rear and front cameras on mobile devices. An end-to-end deep convolutional neural network (CNN) structure is designed to process images on symmetric views and predict SH coefficients. We introduce a novel Render Loss to improve the rendering quality of the predicted illumination. A high quality high dynamic range (HDR) panoramic image dataset was developed for training and evaluation. Experiments show that our model produces visually and quantitatively superior results compared to the state-of-the-arts. Moreover, our method is practical for mobile-based applications.

CCS Concepts

•***Human-centered computing*** → ***Mixed / augmented reality;*** •***Computing methodologies*** → ***Scene understanding;*** *Rendering;*

1. Introduction

Recovering illumination information for 3D physical world is a fundamental problem of computer vision and graphics. It benefits a wide range of applications such as mobile augmented reality [OERF*16], movie post-production [VIC17], virtual advertising [WIK17], and virtual clothes trying [RKS*14, ZSZ*12], which all require composing virtual objects into physical scenes. However, these applications are far from delivering users satisfactory experiences. The key problem is that virtual objects are not rendered under the same illumination as the physical environment. Illumination recovery using common mobile devices without additional special devices is quite useful in many cases. It can benefit many mobile augmented reality applications. However, since the environment is usually quite complex with vast details, recovering illumination from single/multiple images with limited field of view (FOV) is challenging. On common mobile cameras with horizontal FOV of about 60 degrees, a single photo only covers less than 5% of the full panorama. Without strong priors or constraints, it is severely ill-posed to recover the full environment map and estimate the rest of unknowns from limited input information.

Previous works tried to solve such kind of problem by either using more input information or reducing the number of output. Many traditional methods tried to use multiple views or image collection as input [LM14, HFB*09, SAC*13]. Some other methods associate input images with additional geometric information, *e.g.* depth [BM13, ZCC16] and scene geometry [KHFH11, BM15, LN16]. Recently, deep learning methods have brought huge progress in computer vision, image processing and natural language processing. Deep neural networks have also been adopted to predict illumination from single image for indoor [GSY*17] and outdoor [HGSH*17] scenes and produced promising results. Although the problem has been studied for decades and recent deep learning methods made a great progress, there is still a long way to achieve high quality illumination estimation for real applications.

In this paper, we propose a novel deep learning based method for illumination estimation. Similar to traditional methods, we handle the problem with more input information and simplified output. We use pairwise photos taken simultaneously from the front and rear cameras of mobile devices as input.

Low-frequency spherical harmonics lighting is employed to represent scene illumination.

Inspired by the previous learning based methods, we design an end-to-end CNN model, which directly predicts SH coefficients from input image pairs.

The main contributions of this paper are as follows:

1. To the best of our knowledge, this is the first work to predict illumination from pairwise photos. Since the input pairwise pho-

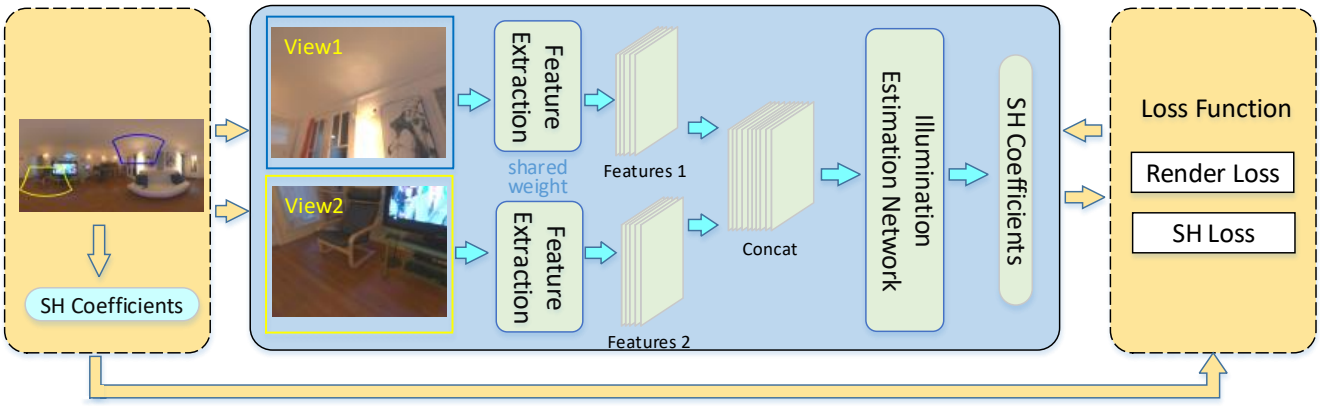

Figure 1: *Overview. In training phase we extract images from an HDR image and compute its corresponding SH coefficient. Then we optimize our network through minimizing the SH loss and render loss between ground truth and the predictions. In testing phase (light-blue box), Pairwise images are feed to the network. After processing by feature extraction network and our fusion network, it produce the illumination represented as SH coefficients.*

tos can be easily accessed from common mobile devices, our method is quite practical for mobile applications.

2. We design a novel CNN architecture for this problem. The model produces high quality illumination predictions. A novel Render Loss is used for training the network. It efficiently performs in-network rendering and optimizing the network parameters with the rendering result.
3. We build a HDR environment dataset with various illumination conditions. It can benefit future research on related topics.

2. Related Work

2.1. Convolutional neural network

Convolutional neural network was first proposed by [LBBH98]. It is powerful tool and has been widely used in many fields with the increasing performance of computer hardware and the publish of very large scale datasets (*e.g.* ImageNet [DDS*09] and ShapeNet [CFG*15]). Various network structures have been proposed to solve different problems in recent years, *e.g.* AlexNet [KSH12], VGG [SZ14], ResNet [HZRS16]. They have achieved great success in many traditional vision problems, such as object detection [GDDM14], classification [KSH12], segmentation [RFB15], etc. Recently, CNNs have also been employed to solve graphics problems, *e.g.* rendering denoising [CKS*17], facial simulation [KAL*17], etc, and have produced promising results. In this paper, inspired by the success in other problems, we employ CNN to predict the illumination of the scene from input images.

2.2. Illumination estimation

The most straightforward way to get accurate illumination distribution in real scenes is capturing. Debevec captures scene illumination by series of photos with a calibration object [Deb98]. Some methods obtain scene illumination with special devices [TKTS11, MRK*13]. While these methods require either complex capturing process or additional devices, some other works focus on estimating illumination from image(s). Most of previous works combine the input image with additional information, such as depth [BM13, ZCC16, SDTC15] and geometry [BM15], or use image collection [LM14, HFB*09, SAC*13]. To further improve the accuracy, some algorithms employ additional preprocessing, *e.g.* depth estimation [KSH*14] scene geometry reconstruction [LXM17] and reflectance estimation [KSH*14, LN16, TY17].

Recently, deep learning has been adopted to the illumination estimation problem. Geoffroy *et al.* [HGSH*17] employed convolutional neural network to predict a parametric illumination model from an outdoor photo. They trained the CNN model on a low dynamic range (LDR) panorama database to predict outdoor illuminations approximated by 6 parameters. Gardner *et al.* [GSY*17] directly recover an HDR environment map from a single LDR photo for indoor scenes. Both of these methods require light source detection which infer the spatial extent and orientation of light sources in LDR panoramic image to build a coarse HDR environment map for training. Compared to these works, we start with two images on opposite views and predict SH lighting coefficients for scene illumination.

2.3. In-network rendering

Rendering is a time-consuming task with complex integral computation. [Goo17] implement a rendering method in a deep learning platforms through Phong shader model. Shu *et al.* [SYH*17] proposed an in-network physically-based face rendering method by representing the light as a 9-dimensional spherical harmonics coefficient vector. Similar to their work, we precalculate SH coefficient maps for rendering 3D objects or environment map. During training, we use them to render images in network. Our in-network rendering can be used to efficiently calculate the Render Loss for training the CNN model.

	View 1	View 2
Input	$224 \times 224 \times 3$	$224 \times 224 \times 3$
Feature Extraction Network	$5 \times 5 \times 256$	$5 \times 5 \times 256$
Concat	$5 \times 5 \times 512$	
Convs 3×3	$5 \times 5 \times 64$	
Dense Layers	2048->1024->512->256->128	
Output	16×3	

Table 1: *Network architecture. Two input images of front and rear views are fed to two branches of the feature extraction network, and we fuse the features and predict 48 SH coefficients(3channels* $\times$ *16). Our network consists of a feature extraction network and a feature fusion network. Feature extraction network is adapted from the pre-trained model for scene classification. Fusion network consists of 5 convolutional layers and 6 fully connected layers. BN, LeakyReLU activation functions are used after each layer (except the output).*

3. Illumination from Pairwise Images

In this work, we aim at directly recovering illumination using a pair of front and rear images of a mobile device as input. We represent global scene illumination with spherical harmonics (SH), and use CNN to regress SH coefficients from input images. Since the input pairwise photos can be easily accessed from common mobile devices, our method is quite practical for mobile applications.

3.1. Spherical harmonics illumination

In computer graphics, illumination can be represented as an environment map or a linear combination of basis functions (*e.g.* spherical harmonics [RH01, Gre03], Haar wavelets [NRH03] or von Mises-Fisher kernels [HNI05]). In this paper, we employ spherical harmonic lighting (SHL) to represent the illumination. SHL is a technique for calculating the lighting on 3D models from area light sources that allows us to capture, relight and display global illumination style images in real time [RH01]. It is widely used in 3D games due to its compactness and high efficiency in rendering. With SH, the lighting distribution can be formulated as:

$$L(s) = \sum_{i=1}^{M} SH_i * y_i(s), \tag{1}$$

where $L(s)$ is the lighting distribution function, s is the spatial direction; M is the number of SH coefficients; y_i is the ith SH basis function; SH_i is the ith SH coefficients. We strongly recommend readers to refer [Gre03] for more details about sphere harmonic lighting. In this paper, we estimate the lighting item L by predicting SH coefficients SH_i from two input images.

3.2. Network structure

A deep convolutional neural network (CNN) is designed to use image pair as input and predict SH coefficients. Fig. 1 illustrates the structure of the proposed network, which contains two main components: the first part extracts features for two input images, and the second one fuses the features and predicts the SH coefficients.

Scene illumination is closely related to the type of the scene, *e.g.* outdoor, indoor, day, night, etc. Therefore, we employ a pre-trained a scene classification network [ZLK*17] to extract image features from both input photos. The extracted symmetric image features are concatenated and followed by convolutional layers(with LeakyReLU) and FC layers to predict final SH coefficients. Table . 1 shows the details of the CNN architecture.

3.3. Loss function

SH Loss. To measure numerical error between predicted SH coefficients and the ground truth, we first use average MSE loss of each SH order as follows:

$$\mathcal{L}_{SH} = \frac{1}{N} \sum_{k=0}^{N-1} \left(\frac{1}{M_k} \sum_{i=0}^{M_k} (SH_{k,i} - \hat{SH}_{k,i})^2 \right), \tag{2}$$

where N is the order of SH, $M_k = 3 * (2k+1)$ is number of the SH coefficients (3 channels) of the k-th order

Although MSE is widely used to evaluate the difference between two vectors, we found it not perform well for optimizing illumination represented by SH coefficients. Low MSE in SH coefficients can not guarantee high quality rendering. Sometimes small difference in SH coefficients can lead to large difference in rendering, while large SH difference can also produce quite similar rendering results (e.g. rotate SH with a small angle). Therefore, we propose a novel *Render Loss* to minimize the difference between rendering results using the ground truth and predicted illuminations.

Render Loss. We render several 3D objects (*e.g.* . sphere, bunny and dragon) into SH coefficient maps. Each SH coefficient map is an image of SH coefficients for each pixel $(width, height, num\_sh\_coeffs)$. Also, environment map can be rendered by such a special SH map. It can be used to efficiently render a color image $(width, height, 3)$ with SH lighting coefficients. During training process, we randomly pick some pre-computed SH maps (including the SH map for rendering environment map) and render them with the ground truth and the predicted SH coefficients respectively:

$$R_{(SH,x,y,c)} = \sum_{i=1}^{M} SHMap(x,y,i) * SH_{c,i}, \tag{3}$$

where M is the number of coefficients, and c is one of the channels of RGB.

Our render loss is defined as the difference (image MSE) between the rendered images using the ground truth and predicted illumination:

$$\mathcal{L}_{render} = \frac{1}{W \times H \times C} \sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{c=1}^{C} (R(SH,x,y,c) - R(\hat{SH},x,y,c))^2, \tag{4}$$

where $R(S,x,y,c)$ is the value of channel c at pixel (x,y) of the rendering result with SH coefficient SH, and W,H is the size of the rendering image. $C = 3$ is the number of channels. The total loss function for training our network is defined as the weighted sum of SH loss and Render loss:

$$\mathcal{L} = w_1 * \mathcal{L}_{SH} + w_2 * \mathcal{L}_{render} \tag{5}$$

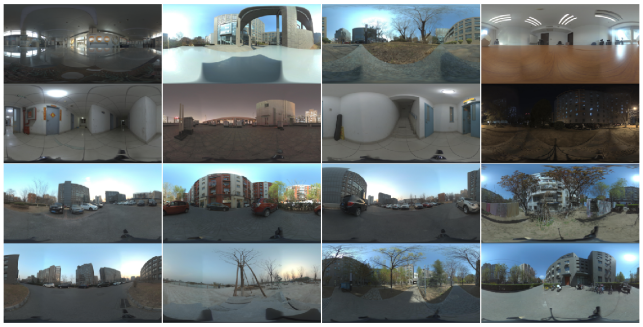

Figure 2: *Examples of HDR panoramas in our HDR dataset.*

where w_1 and w_2 are weights to balance the importance of L_{mse} and L_{render}.

4. Experiments

	GoogLeNet		VGG-16		AlexNet	
	RMSE	DSSIM	RMSE	DSSIM	RMSE	DSSIM
(a)	0.1304	0.0656	0.1269	**0.0642**	0.1638	0.0799
(b)	**0.1292**	**0.0656**	**0.1303**	0.0661	0.1318	0.0717
(c)	0.1329	0.0696	0.1336	0.0718	**0.1239**	**0.0686**

Table 2: *Comparison on testing data using combinations of different backbone models. The performance is evaluated with RMSE and DSSIM metrics. For each backbone, we evaluate the network performance in three ways: (a) Train the entire network from scratch; (b) Fine-tune from the pretrained model. (c) Freeze the pretrained weights of feature extraction network and only train the illumination estimation network; It shows that the optimal way treating to the feature extraction network is different in different network. It depends on which backbone models was adopted and how large the training data is. Based on this comparison results, we use fixed AlexNet as the backbone as the optimal configuration.*

4.1. Dataset

We collected 278 HDR envmaps from online sources. We also captured more than 200 high quality HDR panoramic images (Fig. 2 The size of this data set is still growing. After the publication of the paper, we will release this data to help others work on further research upon this topic.) To build the training and evaluation data for our model, we randomly sample a direction (θ,ϕ) and its opposite directions(-θ,-ϕ) to extract a limited filed-of-view from HDR. In real world, cameras are not always perfectly aligned, thus, we use a Gaussian disturbance of 5° both in the vertical and horizontal direction to the samples. We also scale the pixel values by 2^e, where e is a random value between -1.5 and 1.5, to generate images/SH coefficients with different level of exposures.

We have a total number of 450 HDR panoramic images. For each panorama, we extract images and compute their corresponding SH coefficients and filter out overexposed or underexposed images. Then we split these input/output pairs into three parts for training(80%), validation(10%) and test(10%). We also enforce that data extracted from the same HDR panoramic image do not appear in different subsets.

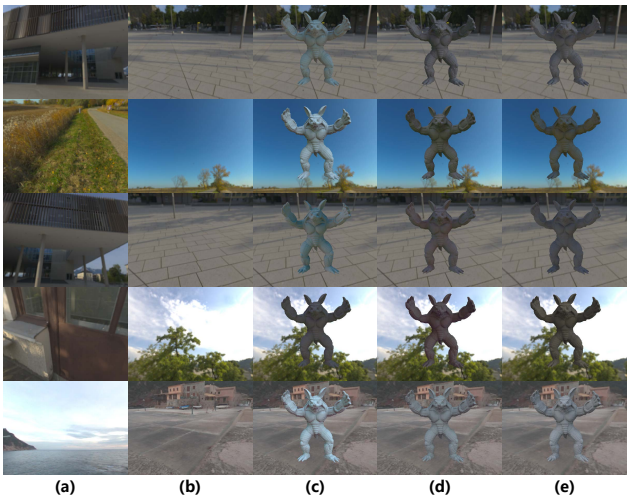

Figure 3: *Results for different ways of treating feature extraction network. (a) Input view-1; (b) Input view-2; (c) Train from scratch; (d) Freeze the pretrained weights of feature extraction network and only train the illumination network;(e) Ground truth.*

Illumination represented as 3-order SH coefficients could produce the result with little error on Lambert surface [RH01]. In this paper, inspired from the strong learning ability of deep CNN, we employ 4-order SH which consist of 48 coefficients for 3 channels to include some high frequency details.

4.2. Network evaluation

We conduct a series of experiment to find the best performance of illumination estimation from pairwise images. For testing our network, we feed images extracted from test dataset to our trained network to produce SH coefficients. Then we render some object using these coefficients, and compute RMSE and DSSIM between these rendering images and the ground truth which are rendered directly using original HDR envirnoment map. It can be ensured that object used for testing are all different from those for computing render loss during training.

4.2.1. Feature extraction networks

In order to find the most suitable backbone model for this problem, we evaluate three models for feature extraction network: VGG-16 [SZ14], AlexNet [KSH12] and GoogLeNet [SLJ*15]. These models were pre-trained for scene classification [ZLK*17] which is closely related to our illumination estimation problems. Table 2 shows the quantitative results on testing data. For the same pre-trained model, we evaluate the performance with: (a) Train the entire network from scratch;(b) Use the pretrained weights and fine-tune from it. (c) Freeze the pre-trained weights of feature extraction network and only train the illumination estimation network; It can be observed that the fixed AlexNet models pre-trained on the

(w_1, w_2)	(0.0, 1.0)		(0.2, 0.8)		(0.5, 0.5)		(0.8, 0.2)		(1.0, 0.0)	
	RMSE	DSSIM	RMSE	DSSIM	RMSE	DSSIM	RMSE	DSSIM	RMSE	DSSIM
GoogLeNet	0.1422	0.0742	0.1334	0.0712	0.1445	0.0785	**0.1329**	**0.0696**	0.1376	0.0732
VGG-16	0.1447	0.0749	0.1561	0.0768	0.1587	0.0765	**0.1336**	0.0718	0.1656	**0.0674**
AlexNet	0.1247	0.0678	0.1268	**0.0675**	0.1479	0.0770	**0.1239**	0.0686	0.1267	0.0682

Table 3: *Comparison on testing data using combinations of different loss weights. The performance is evaluated with RMSE and DSSIM metrics. For each backbone, we compare the results using different loss weights w_1 and w_2 defined at Equation 5. We can observe that using the fusion loss could improve the illumination estimation performance. Based on this result, we use $w_1 = 0.8$ and $w_2 = 0.2$ as the configuration of our best model.*

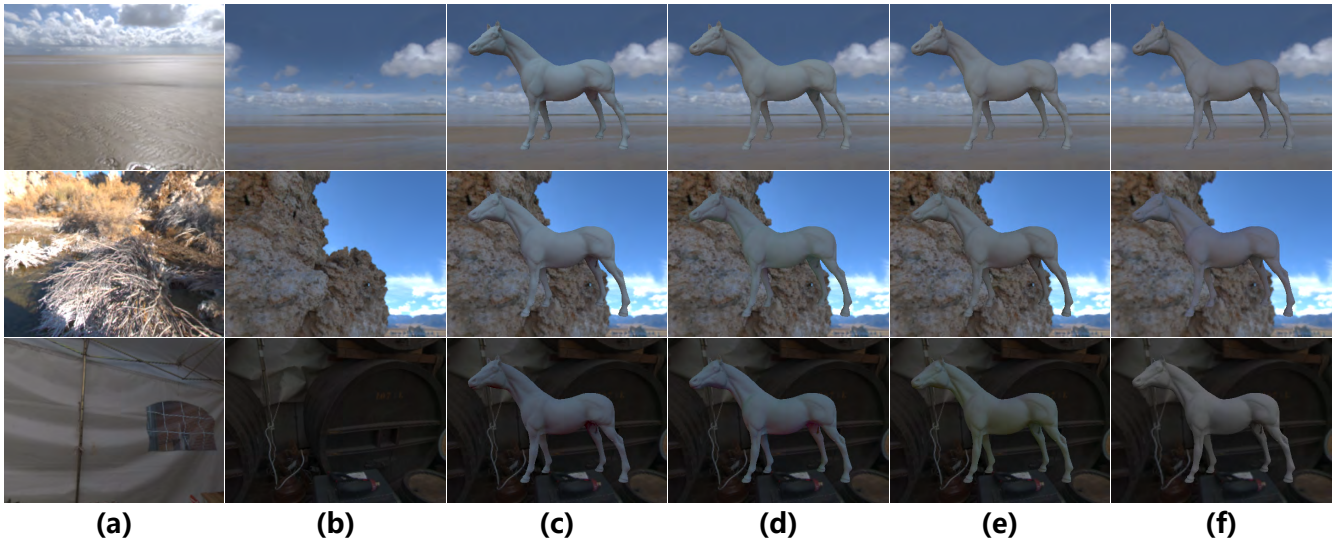

Figure 4: *Evaluation on different combinations of loss functions. (a)(b) Two input views; (c) Only use SH loss (Equation 2); (d) Only use Render Loss (Equation4); (e) Use weighted sum of SH loss and Render loss (Equation 5); (f) Ground Truth.*

data of Places365-Standard [ZLK*17] for scene classification outperforms other settings. Thus, we use AlexNet feature extraction network for in our work. Fig 3 illustrates some visual results.

4.2.2. Loss functions

To evaluate the effect of the render loss and find a optimal loss weight between SH loss and render loss, we conduct comparisons using different combinations of w_1 and w_2.

Fig 4 shows some rendered 3D objects with different illumination estimations which are training with different loss functions. Table 3 shows the quantitative result. We can observe that our method with $w_1 = 0.8$ and $w_2 = 0.2$ can give the more realistic rendering results to ground truth, which shows that both SH loss and render losses can improve the illumination estimation performance.

4.2.3. Fusion layer

The feature extraction network produces two groups of feature map for front and rear view images. Considering the symmetry properties between the two views, they can be fused by either concatenation or subtraction. We compared these two fusion methods and Table 4 shows the numeric evaluations. It can be observed that concatenation is better than subtraction.

	RMSE		DSSIM	
Backbone	Concat	Sub.	Concat	Sub.
GoogLenet	**0.1329**	0.1390	**0.0696**	0.0729
VGG-16	**0.1336**	0.1399	**0.0718**	0.0719
AlexNet	**0.1239**	0.1262	**0.0686**	0.0690

Table 4: *Comparison using different fusion layer for the convolutional feature maps of front and rear images. We use RMSE and DSSIM for evaluation. We find that the concatenation fusion is consistently better.*

4.3. Comparison with previous works

We compared our method with following previous works. Hold-Geoffroy's work [HGSH*17] which is the state-of-art method to predict illumination from outdoor image. GARDNER M.-A's work [GSY*17] which is the state-of-art method to predict indoor illumination from a single image. For outdoor images, we compare our result with [HGSH*17]. For indoor images, we compare the result with [GSY*17]. We predict illumination from images in test dataset by these methods, and render a novel object using these illuminations. Table 5 shows the quantitative result. It shows that

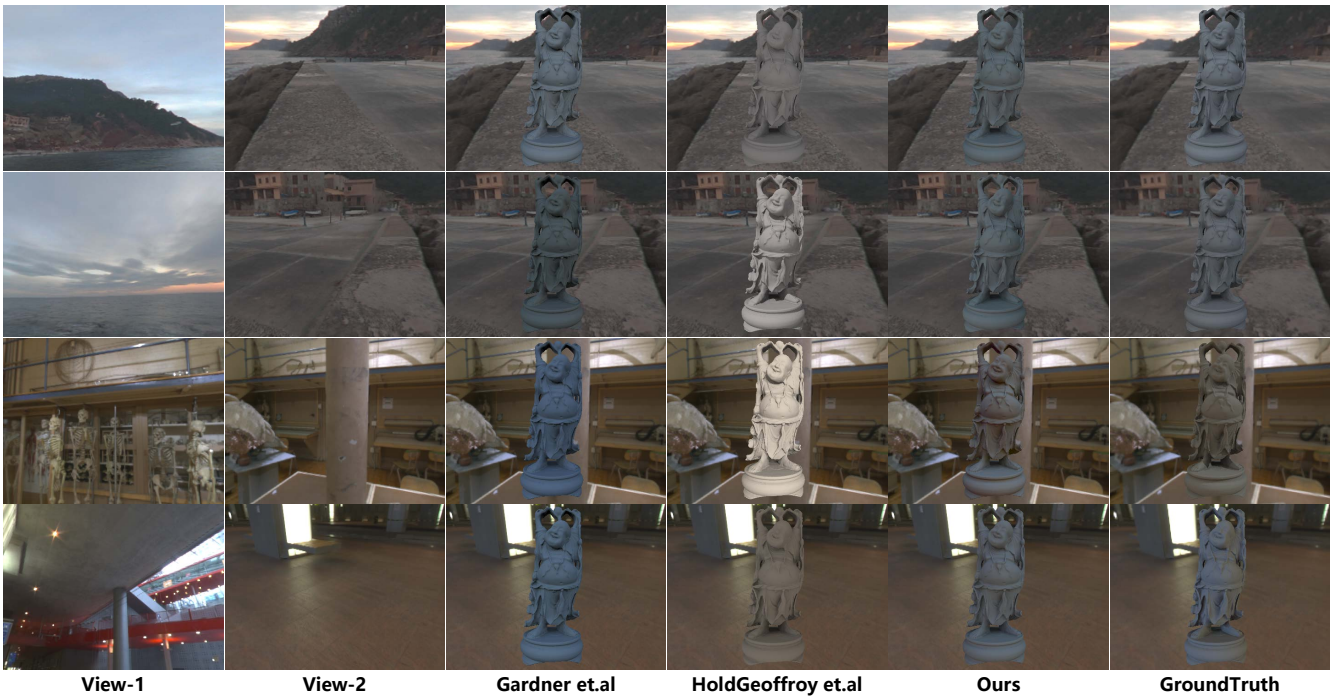

Figure 5: *Comparison with previous work. We compare our results with [GSY*17] and [HGSH*17]. It can be observed that our result is closer to the ground truth, both for indoor and outdoor images.*

Image	Metric	Ours	[HGSH]	[GSY]
Indoor	RMSE	**0.1437**	0.1676	0.2182
	DSSIM	**0.0729**	0.0759	0.1065
Outdoor	RMSE	**0.1185**	0.1609	0.1984
	DSSIM	**0.0670**	0.0780	0.0985
Total	RMSE	**0.1239**	0.1622	0.2027
	DSSIM	**0.0686**	0.0776	0.1003

Table 5: *Quantitative results compared with state-of-the-art methods. [HGSH*17] and [GSY*17] are suitable for outdoor and indoor images respectively. It show that our model produces quantitatively superior results compared to the two state-of-the-arts methods. In addition we noticed that our model performed better when dealing with outdoor scenes. This is mainly due to the small proportion of indoor scenes in the data set.*

our model achieves the best performance. Fig 5 are some examples of comparison.

4.4. Evaluation on mobile camera

We predict illumination using real photos captured by mobile devices, and compare the rendering results with the ground truth. To predict the illumination, we resize the input images to the resolution of 224×224, and feed them to our network. HDR panorama images are captured using a Mi Sphere camera [Mi17]. Each capture is multi-exposed and is fully HDR. Note that HDR capturing device and mobile phone are placed at the same place under the same lighting condition. Fig 6 shows the visual results.

Human face in front camera. When holding mobile phones in front of human body, the face can appear in rear camera view. In our method, we do not handle the face, and the workaround is to ask users to move their head or mobile phone with a small offset. Recently, [CLG*18, YZTL18] try to solve illumination with human face in photos. Human face, in their works, is used as a reference. With a lot of priors on human face, they can achieve some good result. Although we are difference from these methods, their methods can be easily integrated into our framework.

Performance evaluation. In our experiments, our model costs average 0.0391s on a desktop GPU (NVIDIA GTX 1080) and 0.3039s on a single CPU (Intel core i7-6800k) to predict SH coefficients. Our method is designed for mobile applications. Although it cannot achieve real time performance in current phase, we believe with optimization, it can be interactive. And with the advancing of mobile processor, it is promising to meet the performance for mobile applications.

Limitations. Although our model is trained on various illumination images that can produce generally high quality results, there are also some failure cases. Some cases are caused by the lacking of light information in the input image pairs. As Fig 7 shows there is a strong light source at the top of the scene. But the two input views contain few lighting and shadow information. Thus, it is difficult for our method to predict this strong light from them. Another limitation of our method is that Non-Lambert object and illumination

Figure 6: *Visual results using real image pairs captured by mobile cameras. (a) Preview of the HDR environment map; (b) An image of front camera; (c) An image of rear camera; (d) Our Prediction; (e) Ground truth rendering image from HDR envmap. We manually align the extracted ground truth image from HDR image to the image captured by mobile cameras. This leads to a small displacement between them. We scale the input images in linear space to let them have the same exposure value with the ground truth.*

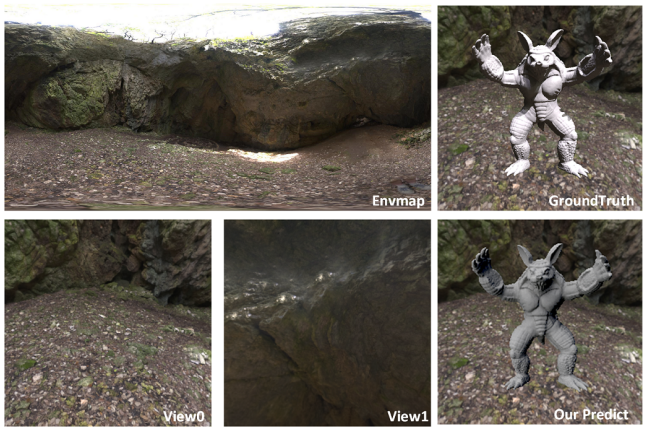

Figure 7: *Failure case caused by the insufficient lighting information contained in the input image pair. There is a strong light source at the top of the scene (left-top). But the image pairs (left-bottom) contain only few information of the light. Thus, our method (right-bottom) can not handle this strong light well, which leads to a result with lower brightness than ground truth (right-top).*

with very high frequency information are unfriendly to our method, since we use SH coefficients to represent the illumination.

5. Conclusion

In this paper, we present a novel method to recover low-frequency scene illumination from an image pair with opposite views. The input images can be easily accessed from rear and front cameras on common mobile phones. We design an efficient CNN model consists of a pre-trained feature extraction network and an illumination estimation network. A Render Loss is employed to train our CNN. Moreover, a data set with high quality HDR panoramic images is developed. The size of the data has reached 500 and is still growing rapidly. We will release this data to help others work on further research upon this topic. Experiments show that our model produces visually and quantitatively superior results to the state-of-the-arts. Moreover, especially designed for mobile phones, our method can make many mobile based augmented reality applications more practical.

Demo, code, model and dataset can be found at our project page: http://cg.tuoo.me/project/shlight

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 61620106003, 61331018 and 61473276)

References

[BM13] BARRON J. T., MALIK J.: Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 17–24. 1, 2

[BM15] BARRON J. T., MALIK J.: Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence 37*, 8 (2015), 1670–1687. 1, 2

[CFG*15] CHANG A. X., FUNKHOUSER T., GUIBAS L., HANRAHAN P., HUANG Q., LI Z., SAVARESE S., SAVVA M., SONG S., SU H., ET AL.: Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015). 2

[CKS*17] CHAITANYA C. R. A., KAPLANYAN A. S., SCHIED C., SALVI M., LEFOHN A., NOWROUZEZAHRAI D., AILA T.: Interactive reconstruction of monte carlo image sequences using a recurrent denoising autoencoder. *ACM Transactions on Graphics (TOG) 36*, 4 (2017), 98. 2

[CLG*18] CALIAN D. A., LALONDE J.-F., GOTARDO P., SIMON T., MATTHEWS I., MITCHELL K.: From faces to outdoor light probes. In *Computer Graphics Forum* (2018), vol. 37, Wiley Online Library, pp. 51–61. 6

[DDS*09] DENG J., DONG W., SOCHER R., LI L.-J., LI K., FEI-FEI L.: Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (2009), IEEE, pp. 248–255. 2

[Deb98] DEBEVEC P.: 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* (1998), ACM, pp. 189–198. 2

[GDDM14] GIRSHICK R., DONAHUE J., DARRELL T., MALIK J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014), pp. 580–587. 2

[Goo17] GOOGLE: Tf-mesh-renderer. https://github.com/google/tf_mesh_renderer, 2017. 2

[Gre03] GREEN R.: Spherical harmonic lighting: The gritty details. In *Archives of the Game Developers Conference* (2003), vol. 56, p. 4. 3

[GSY*17] GARDNER M.-A., SUNKAVALLI K., YUMER E., SHEN X., GAMBARETTO E., GAGNÉ C., LALONDE J.-F.: Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (SIGGRAPH Asia) 9*, 4 (2017). 1, 2, 5, 6

[HFB*09] HABER T., FUCHS C., BEKAER P., SEIDEL H.-P., GOESELE M., LENSCH H. P.: Relighting objects from image collections. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (2009), IEEE, pp. 627–634. 1, 2

[HGSH*17] HOLD-GEOFFROY Y., SUNKAVALLI K., HADAP S., GAMBARETTO E., LALONDE J.-F.: Deep outdoor illumination estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition* (2017). 1, 2, 5, 6

[HNI05] HARA K., NISHINO K., IKEUCHI K.: Multiple light sources and reflectance property estimation based on a mixture of spherical distributions. In *Tenth IEEE International Conference on Computer Vision* (2005), pp. 1627–1634. 3

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778. 2

[KAL*17] KARRAS T., AILA T., LAINE S., HERVA A., LEHTINEN J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG) 36*, 4 (2017), 94. 2

[KHFH11] KARSCH K., HEDAU V., FORSYTH D., HOIEM D.: Rendering synthetic objects into legacy photographs. In *ACM Transactions on Graphics (TOG)* (2011), vol. 30, ACM, p. 157. 1

[KSH12] KRIZHEVSKY A., SUTSKEVER I., HINTON G. E.: Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, Pereira F., Burges C. J. C., Bottou L., Weinberger K. Q., (Eds.). Curran Associates, Inc., 2012, pp. 1097–1105. 2, 4

[KSH*14] KARSCH K., SUNKAVALLI K., HADAP S., CARR N., JIN H., FONTE R., SITTIG M., FORSYTH D.: Automatic scene inference for 3d object compositing. *ACM Transactions on Graphics (TOG) 33*, 3 (2014), 32. 2

[LBBH98] LECUN Y., BOTTOU L., BENGIO Y., HAFFNER P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324. 2

[LM14] LALONDE J.-F., MATTHEWS I.: Lighting estimation in outdoor image collections. In *3D Vision (3DV), 2014 2nd International Conference on* (2014), vol. 1, IEEE, pp. 131–138. 1, 2

[LN16] LOMBARDI S., NISHINO K.: Reflectance and illumination recovery in the wild. *IEEE transactions on pattern analysis and machine intelligence 38*, 1 (2016), 129–141. 1, 2

[LXM17] LIU B., XU K., MARTIN R. R.: Static scene illumination estimation from videos with applications. *Journal of Computer Science and Technology 32*, 3 (2017), 430–442. 2

[Mi17] MI: Mi sphere camera kit. https://www.mi.com/us/mi-sphere-camera-kit/, 2017. 6

[MRK*13] MANAKOV A., RESTREPO J., KLEHM O., HEGEDUS R., EISEMANN E., SEIDEL H.-P., IHRKE I.: A reconfigurable camera add-on for high dynamic range, multispectral, polarization, and light-field imaging. *ACM Transactions on Graphics 32*, 4 (2013), 47–1. 2

[NRH03] NG R., RAMAMOORTHI R., HANRAHAN P.: All-frequency shadows using non-linear wavelet lighting approximation. *Proc Acm Siggraph 22*, 3 (2003), 376–381. 3

[OERF*16] ORTS-ESCOLANO S., RHEMANN C., FANELLO S., CHANG W., KOWDLE A., DEGTYAREV Y., KIM D., DAVIDSON P. L., KHAMIS S., DOU M., ET AL.: Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (2016), ACM, pp. 741–754. 1

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2015), Springer, pp. 234–241. 2

[RH01] RAMAMOORTHI R., HANRAHAN P.: An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* (2001), ACM, pp. 497–500. 3, 4

[RKS*14] ROGGE L., KLOSE F., STENGEL M., EISEMANN M., MAGNOR M.: Garment replacement in monocular video sequences. *ACM Transactions on Graphics (TOG) 34*, 1 (2014), 6. 1

[SAC*13] SHAN Q., ADAMS R., CURLESS B., FURUKAWA Y., SEITZ S. M.: The visual turing test for scene reconstruction. In *3DTV-Conference, 2013 International Conference on* (2013), IEEE, pp. 25–32. 1, 2

[SDTC15] SHI J., DONG Y., TONG X., CHEN Y.: Efficient intrinsic image decomposition for rgbd images. In *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology* (2015), ACM, pp. 17–25. 2

[SLJ*15] SZEGEDY C., LIU W., JIA Y., SERMANET P., REED S., ANGUELOV D., ERHAN D., VANHOUCKE V., RABINOVICH A.: Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1–9. 4

[SYH*17] SHU Z., YUMER E., HADAP S., SUNKAVALLI K., SHECHTMAN E., SAMARAS D.: Neural face editing with intrinsic image disentangling. 5444–5453. 2

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556* (2014). 2, 4

[TKTS11] TOCCI M. D., KISER C., TOCCI N., SEN P.: A versatile hdr video production system. *ACM Transactions on Graphics (TOG) 30*, 4 (2011), 41. 2

[TY17] TODO H., YAMAGUCHI Y.: Estimating reflectance and shape of objects from a single cartoon-shaded image. *Computational Visual Media 3*, 1 (2017), 21–31. 2

[VIC17] VICON: Boujou. https://www.vicon.com/products/software/boujou, 2017. 1

[WIK17] WIKI: Virtualadvertising. https://en.wikipedia.org/wiki/Virtual_advertising, 2017. 1

[YZTL18] YI R., ZHU C., TAN P., LIN S.: Faces as lighting probes via unsupervised deep highlight extraction. *arXiv preprint arXiv:1803.06340* (2018). 6

[ZCC16] ZHANG E., COHEN M. F., CURLESS B.: Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics (TOG) 35*, 6 (2016), 174. 1, 2

[ZLK*17] ZHOU B., LAPEDRIZA A., KHOSLA A., OLIVA A., TORRALBA A.: Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017). 3, 4, 5

[ZSZ*12] ZHOU Z., SHU B., ZHUO S., DENG X., TAN P., LIN S.: Image-based clothes animation for virtual fitting. In *SIGGRAPH Asia* (2012), pp. 33:1–33:4. 1