Using Extended Light Sources for Relighting from a Small Number of Images

Toshiki Hirao, Ryo Kawahara[a] and Takahiro Okabe[b]

Department of Artificial Intelligence, Kyushu Institute of Technology,
680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

Keywords: Relighting, Display-Camera System, Specular Reflection, Extended Light Sources, End-fo-End Optimization.

Abstract: Relighting real scenes/objects is useful for applications such as augmented reality and mixed reality. In general, relighting of glossy objects requires a large number of images, because specular reflection components are sensitive to light source positions/directions, and then the linear interpolation with sparse light sources does not work well. In this paper, we make use of not only point light sources but also extended light sources for efficiently capturing specular reflection components and achieve relighting from a small number of images. Specifically, we propose a CNN-based method that simultaneously learns the *illumination module* (illumination condition), *i.e.* the linear combinations of the point light sources and the extended light sources under which a small number of input images are taken and the *reconstruction module* which recovers the images under arbitrary point light sources from the captured images in an end-to-end manner. We conduct a number of experiments using real images captured with a display-camera system, and confirm the effectiveness of our proposed method.

1 INTRODUCTION

Synthesizing photo-realistic images of a scene/an object under arbitrary illumination environment is one of the most important issues in the interdisciplinary field between computer vision and computer graphics. An approach to synthesizing such images from the real images of the scene/object taken under various lighting conditions is called image-based rendering or (image-based) *relighting* in particular. Relighting real scenes/objects is useful for applications such as augmented reality and mixed reality (Debevec, 1998; Sato et al., 1999; Debevec et al., 2000). In this study, we focus on relighting from the images captured with a display (Schechner et al., 2003; Peers et al., 2009), but our proposed method could be extended to relighting with a light stage (Debevec et al., 2000; Wenger et al., 2003; Hawkins et al., 2004; Wenger et al., 2005; Einarsson et al., 2006; Fuchs et al., 2007; Ghosh et al., 2011).

According to the superposition principle, an image of an object taken under two light sources is a linear combination (convex combination in a strict sense) of the two images, each of which is captured under one of the light sources. Therefore, we can synthesize the image of an object under arbitrary illumination environment by combining the real images of the object taken in advance under various lighting conditions, *e.g.* various light source positions on a display.

In general, the reflected light on an object surface consists of a diffuse reflection component and a specular reflection component. It is known that the image of a Lambertian object under a novel light source direction is represented by the linear combination of the three images of the object taken under non-coplanar light source directions (Shashua, 1997). In other words, the diffuse reflection component observed at a surface point under a novel light source is given by interpolating those under three known light sources. Therefore, we can achieve relighting of Lambertian objects from a small number of images taken under sparse light sources, *e.g.* sparse positions on a display.

On the other hand, relighting of glossy objects is still an open problem to be addressed. This is because specular reflection components are sharp and sensitive to light source positions/directions, and then the linear interpolation with sparse light sources does not work well. Therefore, it requires a large number of

[a] https://orcid.org/0000-0002-9819-3634
[b] https://orcid.org/0000-0002-2183-7112

Hirao, T., Kawahara, R. and Okabe, T.
Using Extended Light Sources for Relighting from a Small Number of Images.
DOI: 10.5220/0012421500003660
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 607-615
ISBN: 978-989-758-679-8; ISSN: 2184-4321

images taken under dense light sources; the smoother a surface is, the larger the number of required images is. Unfortunately, the use of denser light source positions makes the required capture time longer.

Accordingly, in this paper, we make use of not only point light sources but also *extended light sources*[1] for efficiently capturing specular reflection components and achieve relighting from a small number of images. Specifically, we propose a network that simultaneously learns the *illumination module* (illumination condition), *i.e.* the linear combinations of the point light sources and the extended light sources with various sizes under which a small number of input images are taken and the *reconstruction module* which recovers the images under arbitrary point light sources from the captured images. In other words, we optimize both the illumination module and the reconstruction module in an end-to-end manner, and then achieve relighting from a small number of images.

Especially, we focus on the fact that the illumination condition can be represented by (1×1) convolution kernels, and then simultaneously optimize the illumination module and the reconstruction module in the framework of convolutional neural network (CNN). We conduct a number of experiments using real images captured with a display-camera system, and confirm the effectiveness of our proposed method.

The main contributions of this paper are threefold. First, we propose a novel approach that exploits extended light sources for relighting from a small number of images. Second, we propose a data-driven method using the framework of CNN that simultaneously learns the illumination module and the reconstruction module in an end-to-end manner. Third, we experimentally confirm the effectiveness of our proposed method, in particular the use of extended light sources and the end-to-end optimization.

2 RELATED WORK

2.1 Physics-Based Relighting

Image-based rendering under arbitrary illumination environment is called (image-based) relighting. Based on the superposition principle, Debevec *et al.* (Debevec et al., 2000) propose relighting from the 2,048 real images captured by using a light stage, and then extend their method in speed (Hawkins et al., 2004; Wenger et al., 2005), spectra (Wenger et al., 2003), scale (Einarsson et al., 2006), and model acquisition (Ghosh et al., 2011). Their methods work well for human faces, but denser light sources are required for relighting smoother surfaces in general. Fuchs *et al.* (Fuchs et al., 2007) propose a method for reconstructing the images of an object under dense light sources from those under sparse light sources. Their method interpolates the high-frequency components such as specular reflection components under sparse light source directions via optical flow. Since it implicitly assumes surfaces with smooth BRDFs and normals, the applicability of their method is limited.

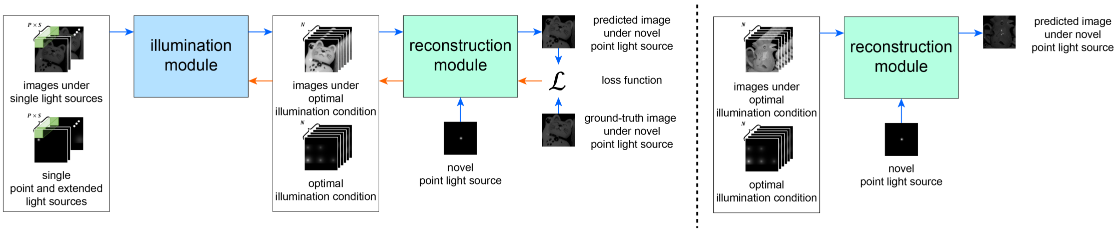
The number of required images for relighting can be reduced by combining image-based rendering with specific physics-based model. For diffuse reflection components, Shashua (Shashua, 1997) shows that the image of a Lambertian object under a novel light source direction is represented by the linear combination of the three images of the object taken under non-coplanar light source directions. For specular reflection components, Lin and Lee (Lin and Lee, 1999) shows that the specular reflection components can be linearly interpolated in the log domain. However, their method implicitly assumes that the specular highlights observed under different light source directions overlap each other, *i.e.* it is applicable to rough surfaces or dense light source directions.

2.2 Learning-Based Relighting

Recently, learning-based methods are proposed for relighting. For human faces, we can exploit the datasets of the face images captured with light stages. Sun *et al.* (Sun et al., 2019) and Zhou *et al.* (Zhou et al., 2019) propose methods for face relighting from a single portrait image. Meka *et al.* (Meka *et al.*, 2019) propose a network that predicts the full 4D reflectance fields of a face from two images captured under spherical gradient illumination, and achieve relighting of non-static faces. Those existing methods work well for face images, but it is not clear whether they are applicable to general classes of objects other than faces.

For general objects, Ren *et al.* (Ren et al., 2015) propose a deep network that models light transport as a non-linear function of light source position and pixel coordinates, and achieve relighting from a relatively small number of images. Xu *et al.* (Xu et al., 2018) achieve image-based relighting from only five directional light sources by jointly learning both the optimal input light directions and the relighting function. Xu *et al.* (Xu et al., 2019) extend their method to image-based rendering under arbitrary lighting and

[1] Note that extended light sources are used in the existing methods (Nayar et al., 1990; Sato et al., 2005), but their purposes are different from ours; the former conducts shape recovery of glossy objects and the latter conducts relighting under low-frequency illumination environment.



(a) network for training phase

(b) network for test phase

Figure 1: Our proposed network with the illumination module and the reconstruction module. The input of the illumination module is point and extended light sources and the images taken under those light sources, and its output is the optimal illumination condition and the images under the optimal illumination condition. The input to the reconstruction module is the output from the illumination module and a novel point light source, and its output is the predicted image under the novel point light source. (a) In the training phase, the illumination module and the reconstruction module are trained on the basis of the loss function $\mathcal{L}$ in an end-to-end manner. (b) In the test phase, we actually capture the images under the trained optimal illumination condition, and then recover the images under novel point light sources by using the trained reconstruction module.

viewing directions. Their method achieve relighting from a small number of images, but there is still room for improvement by using extended light sources. The objective of our study is to investigate the effects of extended light sources for image-based relighting.

2.3 Deep Optics/Sensing

Recently, a number of deep networks that optimize not only application modules but also imaging modules in an end-to-end manner have been proposed. This approach is called *deep optics* or *deep sensing*. A seminal work by Chakrabarti (Chakrabarti, 2016) optimizes the color filter array as well as the demosaicing algorithm in an end-to-end manner. Followed by it, the idea of end-to-end optimization of the imaging modules and the application modules is used for hyperspectral reconstruction (Nie et al., 2018), compressive video sensing (Yoshida et al., 2018), light field acquisition (Inagaki et al., 2018), passive single-view depth estimation (Wu et al., 2019), single-shot high-dynamic-range imaging (Metzler et al., 2020; Sun et al., 2020), seeing through obstructions (Shi et al., 2022), privacy-preserving depth estimation (Tasneem et al., 2022), hyperspectral imaging (Li et al., 2023), and time-of-flight imaging (Li et al., 2022).

Our study also belongs to deep optics/sensing. In contrast to most existing methods that optimize the properties of camera/sensor as well as the application modules, our method optimizes the illumination condition as well as the application module.

3 PROPOSED METHOD

3.1 Overview

Our proposed network consists of two modules: the illumination module and the reconstruction module. Figure 1 illustrates the outline of our network. The input of the illumination module is a set of point light sources and extended light sources with various sizes and the images taken under those light sources. The output of the illumination module is the optimal illumination condition, *i.e.* the optimal linear combinations of those light sources and the images under the optimal illumination condition. The input to the reconstruction module is the output from the illumination module and a novel point light source. The output of the reconstruction module is the predicted image under the novel point light source. Note that we represent light sources by using not the positions (Xu et al., 2018) (and sizes) but the 2D intensity maps, because we consider the linear combinations of point and extended light sources.

In the training phase, we train our proposed network in an end-to-end manner by using the ground truth of the images taken under novel point light sources as shown in Figure 1 (a). Then, we obtain the optimal illumination condition and the reconstruction module that recovers the images under novel point light sources from the images taken under the optimal illumination condition.

In the test phase, we make use of the trained optimal illumination condition and the trained reconstruction module as shown in Figure 1 (b). Specifically, we actually capture the images of a scene/an object under the optimal illumination condition, and then recover the images under novel point light sources by using the reconstruction module. The following subsections explain the details of our network.

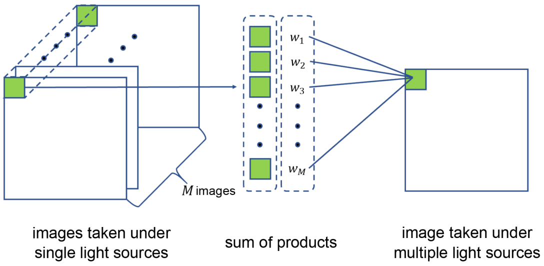

w_1
w_2
w_3
w_M
M images
images taken under single light sources
sum of products
image taken under multiple light sources

Figure 2: The relationship between the superposition principle and a (1×1) convolution kernel. The pixel value of the image taken under multiple light source (right) is represented by the sum of the products between the pixel values of the images taken under single light sources (left) and the intensities of the light sources (the coefficients of the linear combination or the weights). Thus, we can consider the set of the weights as the (1×1) convolution kernel.

3.2 Illumination Module

In general, a display can represent an enormous number of light sources, because its degree of freedom is equal to the number of the display pixels. Here, in order to limit the solution space of the illumination condition, we consider point light sources and extended light sources of Gaussian distributions with $(S-1)$ standard deviations whose centers are at P positions on a display, *i.e.* we consider $P \times S$ light sources in total. The objective of our illumination module is to optimize the N set of intensities of the PS light sources.

We call a set of intensities of the PS light sources a display pattern. According to the superposition principle, we can represent the image captured under a display pattern, *i.e.* a linear combination of PS light sources as the linear combination of the images each of which is captured when turning only one of the PS light sources on. Here, the display pattern and the image captured under the display pattern share the same coefficients of the linear combination w_m $(m = 1, 2, 3, ..., M)$ where $M = PS$.

In order to optimize the illumination condition, we focus on the fact that general display patterns can be represented by (1×1) convolution kernels on the basis of the superposition principle. Specifically, since a display pattern is a linear combination of PS light sources, it is represented by the sum of the products between the pixel values at each pixel of the intensity maps of the PS light sources and the coefficients of the linear combination w_m. It is the same for the image captured under the display pattern. Thus, the weights of the (1×1) convolution kernel correspond to the coefficients of the linear combination w_m as shown in Figure 2.

In our implementation, we represent a general display pattern in two steps. First, for each light source position, we combine the S light sources with the same center position by using a (1×1) convolution kernel as shown in Figure 3 (a), and then normalize the intensities so that the maximal intensity is equal to the maximal pixel value of the display, *e.g.* 255 for an 8-bit display. Second, as shown in Figure 3 (b), we combine those P light source combinations by using a (1×1) convolution kernel and normalize it, and then obtain a display pattern. Thus, our illumination module consists of the two (1×1) convolution layers. Note that when we use N images (and display patterns) for relighting, we use $(P+1)N$ convolution kernels and then optimize $(S+1)PN$ weights in total.

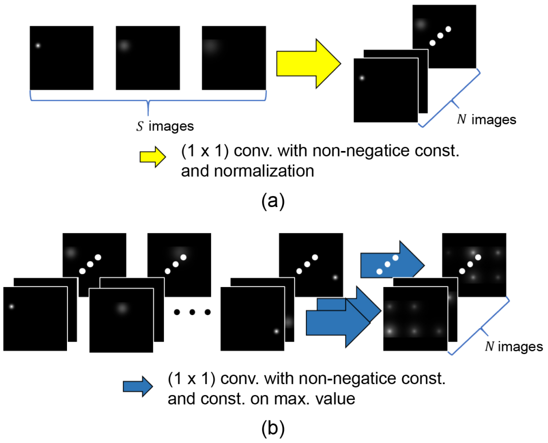

S images
N images
(1 x 1) conv. with non-negatice const. and normalization
(a)
N images
(1 x 1) conv. with non-negatice const. and const. on max. value
(b)

Figure 3: Our illumination module optimizes the illumination condition by representing it in two steps. (a) We combine the S light sources with the same center position by using a (1×1) convolution kernel, and obtain N combinations for each light source position. (b) We combine those N light source combinations for each position by using N (1×1) convolution kernels, and obtain N combinations of point and extended light sources.

Finally, we add two artificial noises to the images under the optimal illumination condition: one obeys Gaussian distribution [2] and the other obeys uniform distribution. The latter is for taking the quantization of a pixel value into consideration. Therefore, the image under darker light sources is more contaminated by the quantization errors of pixel values.

3.3 Reconstruction Module

Note that our substantive proposals are the illumination module and the end-to-end optimization of the illumination module and the reconstruction module.

[2] We use real images which inherently contain noises for training, but random noises are almost canceled out by linearly combining the images. Therefore, we add artificial noises to the linear combination of the real images in order to simulate the noises in one-shot image taken under multiple light sources.

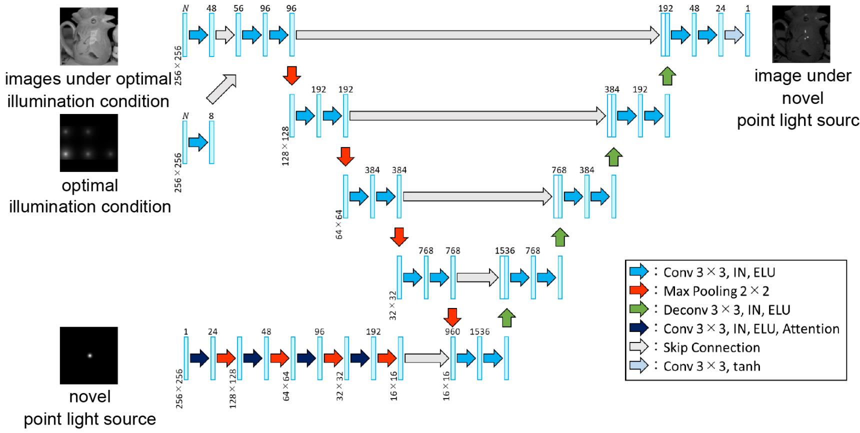


Figure 4: Our reconstruction module; the input to the encoder is the optimal illumination condition and the images under the optimal illumination condition, and the output from the decoder is the image under the novel point light source. The information of a novel point light source is fed at the bottleneck of the U-Net.

Then, we could use an arbitrary end-to-end network for the reconstruction module.

Our current implementation is based on the well-known U-Net architecture (Ronneberger et al., 2015), *i.e.* an encoder-decoder structure with skip connections. It is widely used not only for image-to-image translation (Isola et al., 2017; Liu et al., 2018; Ho et al., 2020; Rombach et al., 2022) but also for deep optics/sensing (Nie et al., 2018; Xu et al., 2018; Wu et al., 2019; Metzler et al., 2020; Sun et al., 2020; Shi et al., 2022). Since deep optics/sensing often adds a kind of illumination module ahead of a conventional application module, the skip connections, which allows information to reach deeper layers and can mitigate the problem of vanishing gradients, are important. Note that the number of feature maps at each layer is optimized by using Optuna (Akiba et al., 2019).

Figure 4 illustrates our reconstruction module. The input to the encoder is the optimal illumination condition and the images under the optimal illumination condition. The sizes of both the illumination condition (2D intensity maps) and the images are 256×256. We repeatedly use the convolution with the kernel size of 3×3, the instance normalization (Ulyanov et al., 2016), the activation function of the ELU (Clevert et al., 2016), and the max pooling with the size of 2×2.

The information of a novel point light source is fed at the bottleneck of the U-Net as an intensity map with 256×256 pixels. We repeatedly use the convolution with the kernel size of 3×3, the instance normalization, the ELU, and the max pooling with the size of 2×2 also for the intensity map of the novel point light source. In addition, we apply the attention mechanism (Xu et al., 2015) for the feature map of the novel point light source. Then, it is merged with the encoded feature maps of the optimal illumination condition and the corresponding images.

The output from the decoder is the image with 256×256 pixels under the novel point light source. We repeatedly use the deconvolution with the kernel size of 3×3, the instance normalization, and the ELU. In addition, the feature maps of each layer of the encoder are used thorough the skip connections. We use the convolution with the kernel size of 3×3 and the activation function of tanh at the last layer, and obtain the image under the novel point light source.

3.4 Optimization

Thus, the illumination condition can be represented as the weights of the convolution kernels, and then we simultaneously learn them as well as the reconstruction module via a CNN-based network in an end-to-end manner. Our proposed network is trained by minimizing the loss function $\mathcal{L}$ of the mean squared errors between the predicted and the ground-truth images under novel point light sources.

4 EXPERIMENTS

4.1 Display-Camera System

As shown in Figure 5, we placed a set of objects on a shelf in front of an LCD, and then captured the images of those objects under varying illumination conditions. We used the LCD as a programmable light source; we realized point light sources and extended light sources with various sizes by display-

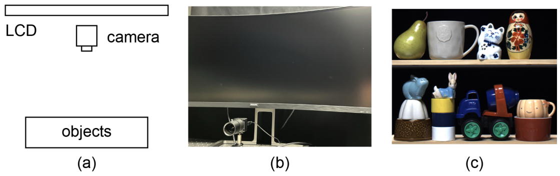


Figure 5: Our display-camera system: (a) the configuration of a display, a camera, and objects, (b) the display and the camera, and (c) the objects on a shelf.

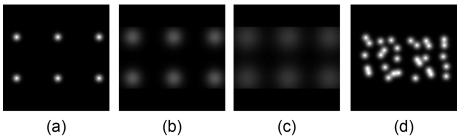


Figure 6: The extended light sources of Gaussian distributions with the standard deviations of (a) 20, (b) 40, and (c) 90 pixels, and (d) point light sources with random positions on the display.

ing the intensity patterns of those light sources on the LCD. We used an LCD of 439P9H1 from Philips and a monochrome camera of CMLN-13S2M-CS from Point Grey. We confirmed that the radiometric response function of the camera is linear, but that of the display is non-linear. The radiometric response function of the display was calibrated by using the set of images captured by varying input pixel values of the display.

4.2 Setup

We captured the images of 15 scenes in total; the images of 9, 3, and 3 scenes were used for training, validation, and test respectively. In order to efficiently train our proposed network from a relatively small number of scenes, the image patches with 256×256 pixels were cropped from each captured image. Therefore, the actual numbers of scenes are considered to be 540, 180, and 108 for training, validation, and test respectively.

As shown in Figure 6, we consider the extended light sources of Gaussian distributions with $S = 3$ standard deviations whose centers are at $P = 6$ positions on the display, *i.e.* we consider $P \times S = 18$ light sources in total. We set the standard deviations of the Gaussian distributions to 20, 40, and 90 pixels for the display area with 950×1800 pixels. We can realize a point light source by turning a single pixel on the display on, but such light source is too dark to illuminate scenes with sufficient intensity. Then, we consider the extended light source with the smallest size as a point light source. In addition, we captured 30 ground-truth images per scene under point light sources with random positions inside the $P(=6)$ positions on the display, and used them for the training and validation. We captured 153 ground-truth images per scene in a similar manner, and then used them for the test.

We used the optimization algorithm of the Adam (Kingma and Ba, 2016) for training. We set the initial learning rate to a relatively large value of 1.0×10^{-3} so that the problem of vanishing gradients at the input and nearby layers is mitigated, and then gradually decreased it. We used the loss function of the MSE, and used the MSE and SSIM for validation. The weights of the illumination module are initialized with the uniform distribution from 0.4 to 0.6, and the other weights are initialized by using the He normal initialization (He et al., 2015). The mean and standard deviation of the Gaussian noises described in Section 3.2 are 0 and 2 for 8-bit images respectively. It took about 27 hours for training our proposed network with 2,700 iterations.

4.3 Results

To confirm the effectiveness of our proposed method, in particular the use of extended light sources as well as the end-to-end optimization of the illumination module and the reconstruction module, we compared the following five
methods:

A **Linear Interpolation with Point Light Sources:** the linear interpolation of the $N_A = P\ (= 6)$ images taken under the P point light sources.

B **Nonlinear interpolation with point light sources:** the nonlinear interpolation of the $N_B = P\ (= 6)$ images taken under the P point light sources. The nonlinear interpolation is trained by using our reconstruction module.

C **Our Method Without the Illumination Module:** our reconstruction module is used for random and fixed combinations of point and extended light sources. N_C images are used for reconstruction.

D **Our Method:** the end-to-end optimization with the illumination module.
N_D images are used for reconstruction.

E **Reconstruction from All Point and Extended Light Sources:** our reconstruction module is trained by using the $N_E = P \times S\ (= 18)$ images taken under each of the PS light sources for reference.

In summary, the numbers of captured images are as follows: $N_A = N_B = P = 6$ by definition, $N_C = N_D = 6$

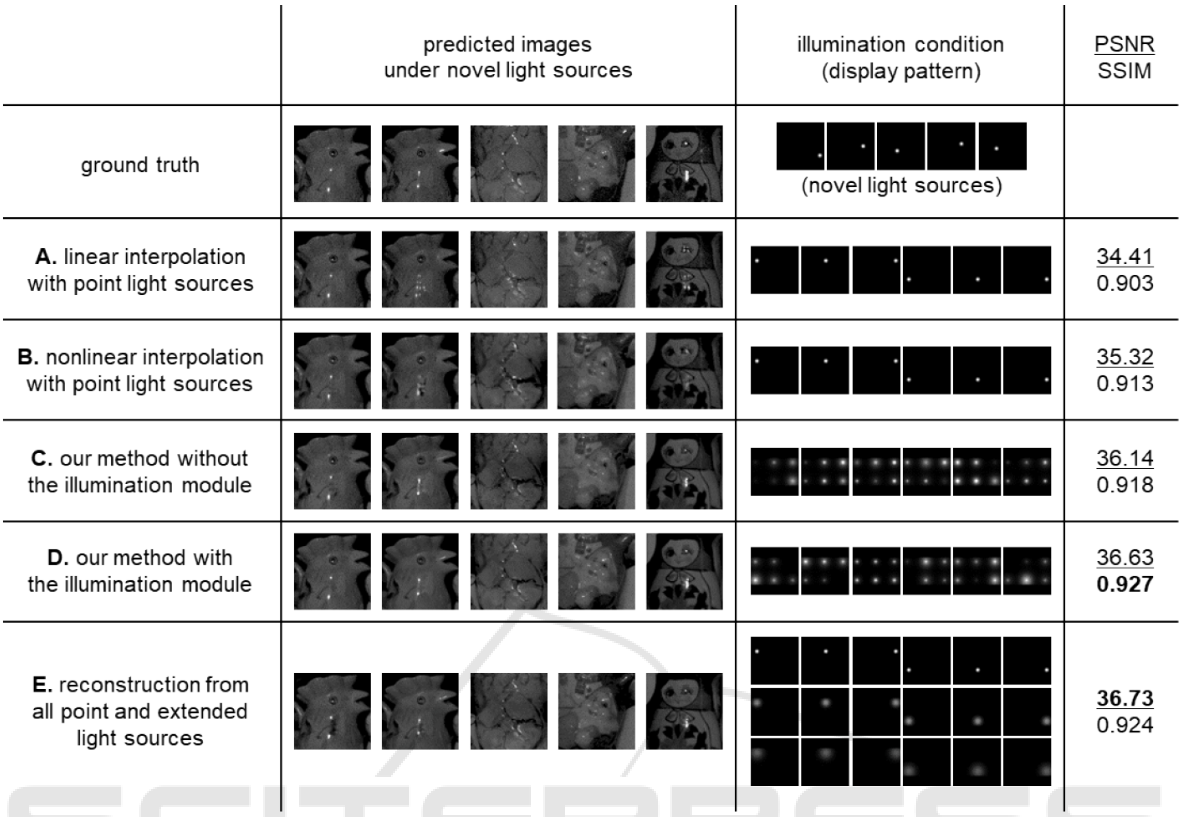

predicted images
under novel light sources
illumination condition
(display pattern)
PSNR
SSIM
ground truth
(novel light sources)
A. linear interpolation with point light sources
34.41
0.903
B. nonlinear interpolation with point light sources
35.32
0.913
C. our method without the illumination module
36.14
0.918
D. our method with the illumination module
36.63
0.927
E. reconstruction from all point and extended light sources
36.73
0.924

Figure 7: The qualitative and quantitative comparison: the predicted images under novel point light sources, the illumination conditions, and the PSNRs and SSIMs from left to right, and the ground-truth images and the results of A through E from top to bottom. We applied the gamma correction to those images only for display purpose.

for comparison with A and B, and $N_E = PS = 18$ [3].

Figure 7 shows the qualitative and quantitative results of those methods: the reconstructed images under novel point light sources, the illumination conditions, and the PSNRs and SSIMs from left to right, and the ground-truth images and the results of A through E from top to bottom. The higher PSNR and SSIM are, the better.

A vs. B: We can compare the performances of the linear interpolation and the nonlinear interpolation with point light sources. We can see qualitatively and quantitatively that the nonlinear interpolation works better than the linear interpolation. In particular, the specular highlight reconstructed by the linear interpolation is just a linear combination of the original specular highlights with the same positions, and then multiple highlights are observed in the reconstructed image although a single highlight is observed in the corresponding area of the ground-truth image.

(A, B) vs. (C, D): We can compare the performances with/without extended light sources. We can see that the methods using extended light sources (C, D) perform better than the methods using only point light sources (A, B) in terms of PSNR and SSIM.

C vs. D: We can compare the performances with/without our illumination module. We can see that the methods using the illumination module (D) outperform the method without the illumination module (C) in terms of PSNR and SSIM.

Therefore, we can conclude that the use of extended light sources is effective from the comparison between (A, B) and (C, D) and that the end-to-end optimization, in particular our illumination module is effective from the comparison between C and D. In addition, we can say that our reconstruction module works well from the comparison between A and B. Note that E works best simply because the number of images is 3 times larger than the other methods, but our method with 6 images works well similarly.

[3] Note that our proposed method is related to light transport acquisition such as multiplexed illumination (Schechner et al., 2003) and compressive sensing (Peers et al., 2009), but the number of required images are far smaller than them.

It is interesting that the optimal illumination conditions themselves show the effectiveness of extended light sources. Specifically, our proposed method uses the combinations of various point and extended light sources.

5 CONCLUSION AND FUTURE WORK

We achieved relighting from a small number of images by using not only point light sources but also extended light sources for efficiently capturing specular reflection components. Specifically, we proposed a CNN-based method that simultaneously learns the illumination module and the reconstruction module in an end-to-end manner. We conducted a number of experiments using real images captured with a display-camera system, and confirmed the effectiveness of our proposed method. The extension of our method for other high-frequency components of images such as cast shadows and caustics is one of the future directions of our study.

ACKNOWLEDGEMENTS

This work was partly supported by JSPS KAKENHI Grant Numbers JP23H04357 and JP20H00612.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., and M.Koyama (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proc. ACM SIGKDD KDD2019*, pages 2623–2631.

Chakrabarti, A. (2016). Learning sensor multiplexing design through back-propagation. pages 3089–3097.

Clevert, D., Unterthiner, T., and Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (ELUs). In *Proc. ICLR2016*.

Debevec, P. (1998). Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Proc. ACM SIGGRAPH1998*, pages 189–198.

Debevec, P., Hawkins, T., Tchou, C., Duiker, H., Sarokin, W., and Sagar, M. (2000). Acquiring the reflectance field of a human face. In *Proc. SIGGRH2000*, pages 145–156.

Einarsson, P., Chabert, C.-F., Jones, A., Ma, W.-C., Lamond, B., Hawkins, T., Bolas, M., Sylwan, S., and Debevec, P. (2006). Relighting human locomotion with flowed reflectance fields. In *Proc. EGSR2006*, pages 183–194.

Fuchs, M., Lensch, H., Blanz, V., and Seidel, H. (2007). Superresolution reflectance fields: Synthesizing images for intermediate light directions. In *Proc. EGSR2007*, volume 26, pages 447–456.

Ghosh, A., Fyffe, G., Tunwattanapong, B., Busch, J., Yu, X., and Debevec, P. (2011). Multiview face capture using polarized spherical gradient illumination. *ACM TOG*, 30(6):1–10.

Hawkins, T., Wenger, A., Tchou, C., Gardner, A., Göransson, F., and Debevec, P. (2004). Animatable facial reflectance fields. In *Proc. EGSR2004*, pages 309–319.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. IEEE ICCV2015*, pages 1026–1034.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Inagaki, Y., Kobayashi, Y., Takahashi, K., Fujii, T., and Nagahara, H. (2018). Learning to capture light fields through a coded aperture camera. In *Proc. ECCV2018*, pages 418–434.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. (2017). Image-to-image translation with conditional adversarial networks. In *Proc. IEEE CVPR2017*, pages 5967–5976.

Kingma, D. and Ba, L. (2016). Adam: A method for stochastic optimization. In *Proc. ICLR2016*.

Li, J., Yue, T., Zhao, S., and Hu, X. (2022). Fisher information guidance for learned time-of-flight imaging. In *Proc. IEEE/CVF CVPR2022*, pages 16313–16322.

Li, K., Dai, D., and Van, G. L. (2023). Jointly learning band selection and filter array design for hyperspectral imaging. In *Proc. IEEE WACV2023*, pages 6384–6394.

Lin, S. and Lee, S. (1999). A representation of specular appearance. volume 2, pages 849–854.

Liu, G., Reda, F., Shih, K., Wang, T.-C., Tao, A., and Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In *Proc. ECCV2018*, pages 85–100.

Meka *et al.*, A. (2019). Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. *ACM TOG*, 38(4):Article No.7.

Metzler, C., Ikoma, H., Peng, Y., and Wetzstein, G. (2020). Deep optics for single-shot high-dynamic-range imaging. In *Proc. IEEE/CVF CVPR2020*, pages 1375–1385.

Nayar, S., Ikeuchi, K., and Kanade, T. (1990). Determining shape and reflectance of hybrid surfaces by photometric sampling. *IEEE Trans. Robotics and Automation*, 6(4):418–431.

Nie, S., Gu, L., Zheng, Y., Lam, A., Ono, N., and Sato, I. (2018). Deeply learned filter response functions for hyperspectral reconstruction. In *Proc. IEEE/CVF CVPR2018*, pages 4767–4776.

Peers, P., Mahajan, D., Lamond, B., Ghosh, A., Matusik, W., Ramamoorthi, R., and Debevec, P. (2009). Compressive light transport sensing. *ACM TOG*, 28(1):Article No.3.

Ren, P., Dong, Y., Lin, S., Tong, X., and Guo, B. (2015). Image based relighting using neural networks. *ACM TOG*, 34(4):1–12.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF CVPR2022*, pages 10684–10695.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Proc. MICCAI2015*, pages 234–241.

Sato, I., Okabe, T., Sato, Y., and Ikeuchi, K. (2005). Using extended light sources for modeling object appearance under varying illumination. In *Proc. IEEE ICCV2005*, pages 325–332.

Sato, I., Sato, Y., and Ikeuchi, K. (1999). Acquiring a radiance distribution to superimpose virtual objects onto a real scene. *IEEE TVCG*, 5(1):1–12.

Schechner, Y., Nayar, S., and Belhumeur, P. (2003). A theory of multiplexed illumination. In *Proc. IEEE ICCV 2003*, pages 808–815.

Shashua, A. (1997). On photometric issues in 3d visual recognition from a single 2d image. *IJCV*, 21(1-2):99–122.

Shi, Z., Bahat, Y., Baek, S.-H., Fu, Q., Amata, H., Li, X., Chakravarthula, P., Heidrich, W., and Heide, F. (2022). Seeing through obstructions with diffractive cloaking. *ACM TOG*, 41(4):1–15.

Sun, Q., Tseng, E., Fu, Q., Heidrich, W., and Heide, F. (2020). Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *Proc. IEEE/CVF CVPR2020*, pages 1386–1396.

Sun, T., Barron, J., Tsai, Y.-T., Xu, Z., Yu, X., Fyffe, G., Rhemann, C., Busch, J., Debevec, P., and Ramamoorthi, R. (2019). Single image portrait relighting. *ACM TOG*, 38(4):1–12.

Tasneem, Z., Milione, G., Tsai, Y.-H., Yu, X., Veeraraghavan, A., Chandraker, M., and Pittaluga, F. (2022). Learning phase mask for privacy-preserving passive depth estimation. In *Proc. ECCV2022*, pages 504–521.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. *arXiv:1607.08022*.

Wenger, A., Gardner, A., Tchou, C., Unger, J., Hawkins, T., and Debevec, P. (2005). Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM TOG*, 24(3):756–764.

Wenger, A., Hawkins, T., and Debevec, P. (2003). Optimizing color matching in a lighting reproduction system for complex subject and illuminant spectra. In *Proc. EGWR2003*, pages 249–259.

Wu, Y., Boominathan, V., Chen, H., Sankaranarayanan, A., and Veeraraghavan, A. (2019). Phasecam3d-learning phase masks for passive single view depth estimation. In *Proc. IEEE ICCP2019*, pages 1–12.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML2015*, pages 2048–2057.

Xu, Z., Bi, S., Sunkavalli, K., Hadap, S., Su, H., and Ramamoorthi, R. (2019). Deep view synthesis from sparse photometric images. *ACM TOG*, 38(4):1–13.

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

Yoshida, M., Torii, A., Okutomi, M., Endo, K., Sugiyama, Y., Taniguchi, R., and Nagahara, H. (2018). Joint optimization for compressive video sensing and reconstruction under hardware constraints. In *Proc. ECCV2018*, pages 634–649.

Zhou, H., Hadap, S., Sunkavalli, K., and Jacobs, D. W. (2019). Deep single-image portrait relighting. In *Proc. IEEE ICCV2019*, pages 7194–7202.