Scalable, Detailed and Mask-free Universal Photometric Stereo Supplementary Material

Satoshi Ikehata
sikehata@nii.ac.jp

Appendix A. Failure Cases and Better Imaging

While the main text discusses only the theoretical aspects of the problem, this section discusses the practical aspects of our method by discussing in more detail the limitations regarding the input acquisition and how to capture better images for our method.

Basics about image acquisition. Our image acquisition process is simple. Prepare a scene and take photos of it under different lighting conditions without moving a camera. The light source can theoretically be either active (*e.g.*, using a hand-held light) or passive (*e.g.*, mounting a camera and an object on the same board and moving them around) as long as sufficient changes in illumination occur. Realistically, the most probable situation may involve a combination of dynamic active lights in a static environment.

Our method has no restrictions on the size of scenes. On the other hand, since the proposed method assumes an orthographic camera, extreme projection distortion is not considered. However, as a common practice, the view directions of a perspective projection camera become more parallel with each other around the central field of view, so using only the central region of a sufficiently high-resolution image is not problematic for practical purposes. Throughout the papers (*i.e.*, main and supplementary), we used either a 45mm or 200mm focal length camera based on the object size to capture 4000x4000 images, of which the central 2048x2048 area was used in the preprocessing as described in the main paper.

Failure cases and possible solutions. We observed two major cases of failure in the course of our experiments as illustrated in Fig. 1. First, the performance drastically degrades if the unmasked region contains areas where no or little illumination change exists because our training data (PS-Mix) contains no cases where the light source condition doesn't change or is very weak from image to image in any regions of the image. For example, when a spotlight light is illuminated on an object, the surface normal recovery could fail if the image contains many areas that are not included in the light diameter. Another common case is that the intensity of the dynamic light source is very weak compared to the static one therefore illumination changes between images are scarce. This tends to occur when the method is applied during the daytime or when trying to recover large scenes of wide depth range.

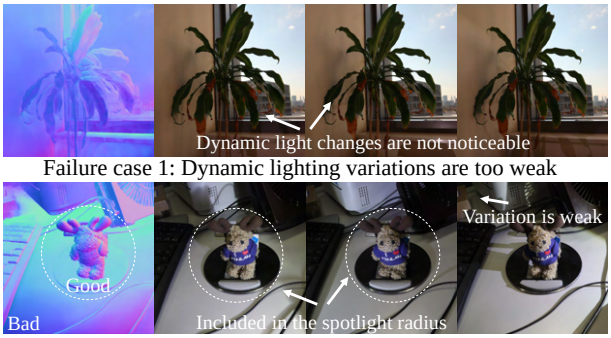


Figure 1. Failure cases. The performance of the proposed method degrades significantly when changes in the illumination environment cannot be observed, whether in part or in the entire image.

There are various possible ways to improve this, such as improving the training data by including such cases (*e.g.*, spot light rendering) or adding a mechanism to identify and ignore regions where light source changes do not occur, but further discussion would go beyond what is allowed in the supplementary, so we leave these issues for the future work.

The better choice of light source. Based on the discussion above, a point light source or surface light source that can illuminate a wide area simultaneously may seem more appropriate, rather than a spotlight that tends to produce areas that are not clearly illuminated. Generally speaking, when automatic exposure control is turned on, the tonal resolution is degraded to increase dynamic range when very dark and bright areas are mixed together. On the other hand, when the entire image is bright, it is possible to represent enough information within a narrow dynamic range, resulting in less image noise. Therefore, to improve the quality of a captured image, it is essential to make the irradiance uniform across the image.

Empirically, we have found that using a ring light for

Figure 2. Our acquisition setup simply needs a movable light source and a camera.

selfies or a smartphone/tablet screen as a light source are the two most effective methods available to us that meet the above conditions. Of these, the selfie-light, which provides sufficient light and is easy to handle, was used in many of the experiments in this paper. The tools used in this work are shown in Fig. 2. Since no calibration of the light source or camera is necessary, all that is required are a single light source, a single camera, and target objects.

When a mask is necessary and when it is not. Basically, the proposed method does not require a mask. As one may have noticed from the paper's results, our method is capable of preserving the depth discontinuities of objects without a mask to a level that is not possible with any existing methods. There are three main factors that make this possible. First, unlike the existing dataset (*i.e.*, PS-Wild), our PS-Mix consists of multiple overlapping objects, and learning is performed without explicitly providing their boundaries. Second, our method is more robust than all existing methods against inter-reflections and cast shadows that occur at depth boundaries. Third, during global interactions of the aggregated features, no local interaction is performed unlike existing methods, so no over-smoothing occurs.

However, there are some cases where the object mask is helpful. The first case is simply when one wishes to recover only the shape of a particular object in a scene. The second case is when we want to explicitly "hide" areas with little lighting variation. During network training, a ground truth mask is always given simultaneously, and the loss function is computed only from pixels in the mask. Consequently, information outside the mask is not taken into account in the prediction during training. In other words, a mask can be used to intentionally hide areas from the network where lighting variations are weak. For this purpose, the mask does not need to follow the contour of the object; a bounding box-like specification is sufficient.

Other points to note on photography. We found that there are other exceptional cases where our method does not work well. If the image correction is too strong, it will fail. For example, recent smartphones apply various image filters to improve the appearance of images after they are taken. As a result, the physically correct shading changes are destroyed. Also, while the proposed method basically does not require HDR (high-dynamic-range) images, it is not as robust with respect to too much over- and under- exposure. Fortunately, the automatic exposure control provided in recent digital cameras and smartphones is very effective to avoid the situations. Similarly to other photometric stereo methods, our method is also helpless with respect to an image blur and an accidental misalignment of images. The above problems can be easily solved by carefully tuning the camera, so they are not critical in practice.

Summary

In conclusion to this section, the following points should be kept in mind when taking photographs.

- To assume an orthographic camera, the object should be placed in the central field of view of a camera with a sufficiently large focal length.
- Ensure that the illumination changes throughout the image. For this purpose, light sources that can illuminate a wide area, such as a point light or a surface light, are better than a spotlight. Alternatively, masks can be used to hide areas of weak illumination variation.
- Turn off software image correction, increase the depth of field to prevent blur, and ensure that the camera does not move while taking photographs.
- The number of images can be small. If you need more, just add more. It only takes a few seconds.

Appendix B. Network Architecture Details

Our entire framework consists of six sub-networks. The scale-invariant spatial-light encoder includes (a) a backbone network for the imagewise feature extraction, (b) a Transformer network for the pixelwise interaction along the light-axis and (c) a feature pyramid network for the fusion of hierarchical feature maps. And in the pixel-sampling Transformer, there are (d) a Transformer network for the feature aggregation along the light-axis and (e) a Transformer network for the feature interaction along the spatial-axis. Finally, we have (f) a MLP for the surface normal prediction. In this section, we detail each network architecture.

Backbone: In our scale-invariant spatial-light encoder, each sub-tensor (*i.e.*, concatenation of a sub-image and a sub-mask) is independently input to ConvNeXt [13] which is a modernized ResNet [6] like architecture taking inspiration from the recent Vision Transformer [5, 12]. The variants of ConvNeXt differ in the number of channels C, and

the number of ConvNeXt blocks B in each stage. We here chose the following configuration.

- ConvNeXt-T: C = (96, 192, 384, 768), B = (3, 3, 9, 3)

The ConvNeXt block includes 7x7 depthwise convolution, 1x1 convolution with the inverted bottleneck design (4x hidden dimension) and 1x1 convolution to undo the hidden dimension. Between convolutions, layer normalization [18] and GeLU [7] activation are placed. The output of ConvNeXt is a stack of feature maps of (B x 96 x R/4 x R/4), (B x 192 x R/8 x R/8), (B x 384 x R/16 x R/16) and(B x 768 x R/32 x R/32) where B is the batch size and R is the input sub-tensor size as defined in the main paper.

Transformer (interaction along light-axis): Given hierarchical feature maps from the backbone network, we pixelwisely apply Transformer [16] to features of individual scales along the light-axis as with [8]. We chose the number of channels in a hidden layer C, and the nuber of Transformer blocks B as follow.

- Transformer: C = (96, 192, 384, 768), B = (0, 1, 2, 4)

The Transformer block projects the input feature to query, key and value vectors whose dimensions are same with the input ones. They are then passed to a multi-head self-attention (the number of heads is 8) with a soft-max and a feed-forward network with two linear layers whose dimensionality of input and output layers is same but one of the inner layer is twice of the input. A residual connection around each of the two sub-layers, followed by layer normalization [18] and dropout ($p = 0.1$).

Feature pyramid network: After the hierarchical feature maps pixelwisely interact with each other using Transformers, feature maps of different scales corresponding to each input image are fused with the feature pyramid network (*i.e.* UPerNet [17]) which was originally proposed for the semantic segmentation task. We simply used an implementation on MMSegmentation [3] without any modifications. The output feature size is (B x R/4 x R/4 x 256).

Transformer (aggregation along light-axis): Given m pixel locations at the input coordinate system, we concatenate each pair of a raw observation and a bilinearly interpolated feature vector from the output of the feature pyramid network to a vector whose dimension is 259 (*i.e.*, 256+3). The feature aggregation network takes K sets of 259-dim feature vectors at the same location as input and perform two Transformer blocks of C=256 (shrunk from 259 to 256 by QKV projection). The output feature is further concatenated with the raw observation and each 259-dim feature vectors are again fed to another three Transformer blocks of C=256. Then, the output K feature vectors are passed to PMA [9] where the number of elements in a set was shrunk from K to one using another Transformer block of C=384.

Transformer (interaction along spatial-axis): At the final step of the pixel-sampling Transformer module, we perform two Transformer blocks (C=384) to communicate features among the m locations. The näive self-attention requires O(m^2) memory consumption, however m (*i.e.*, number of pixel samples) is much larger than K (*i.e.*, number of input images), which makes increasing sample size difficult. Therefore, we instead used the O(m) implementation of the self-attention by [14] to tackle this problem (Note that the computational cost doesn't change).

Normal prediction network: The surface normal predictor is a MLP with one hidden layer whose feature dimension shrank as $384 \rightarrow 192 \rightarrow 3$ and the norm of the output vector is normalized to be a unit surface normal vector at the location.

Appendix C. Reflectance Recovery

As highlighted in the main paper's conclusion, the proposed framework extends beyond surface normal recovery and can be readily applied to surface reflectance recovery by merely substituting the training data and loss function. This versatility enables the proposed method to render the target scene under novel lighting environments, or in other words, achieve novel relighting. However, recovering surface reflectance from images in an uncalibrated setup poses a fundamental ambiguity due to the countless possible combinations of illumination and reflectance, such as a red surface under white light or a white surface under red light. This complexity makes objective evaluation nearly unattainable.

Given the challenges in evaluation, we opted not to feature the results of reflectance recovery in the main paper. Instead, we present them here to demonstrate that our method is not confined to surface normal recovery. Our PS-Mix dataset already incorporates base color, roughness, and metalness maps from AdobeStock [1], which were employed to render images. We simply utilize these maps as training data and train our network using Mean Squared Error (MSE) losses between the predicted and provided base color, roughness, and metalness maps. Once the surface reflectance parameters and surface normals are recovered, we can render images of the scene under novel lighting conditions using any physically-based renderer, such as Blender [2].

Implementation details: We implemented two separate networks for surface normal map recovery and base color, roughness, and metalness maps recovery, respectively. We observed that training a single network for both tasks slightly degraded performance. The network architecture and training methodology were identical to those described in the main paper, with the only difference being the training data and loss functions.

Reflectance representation: The images in both PS-

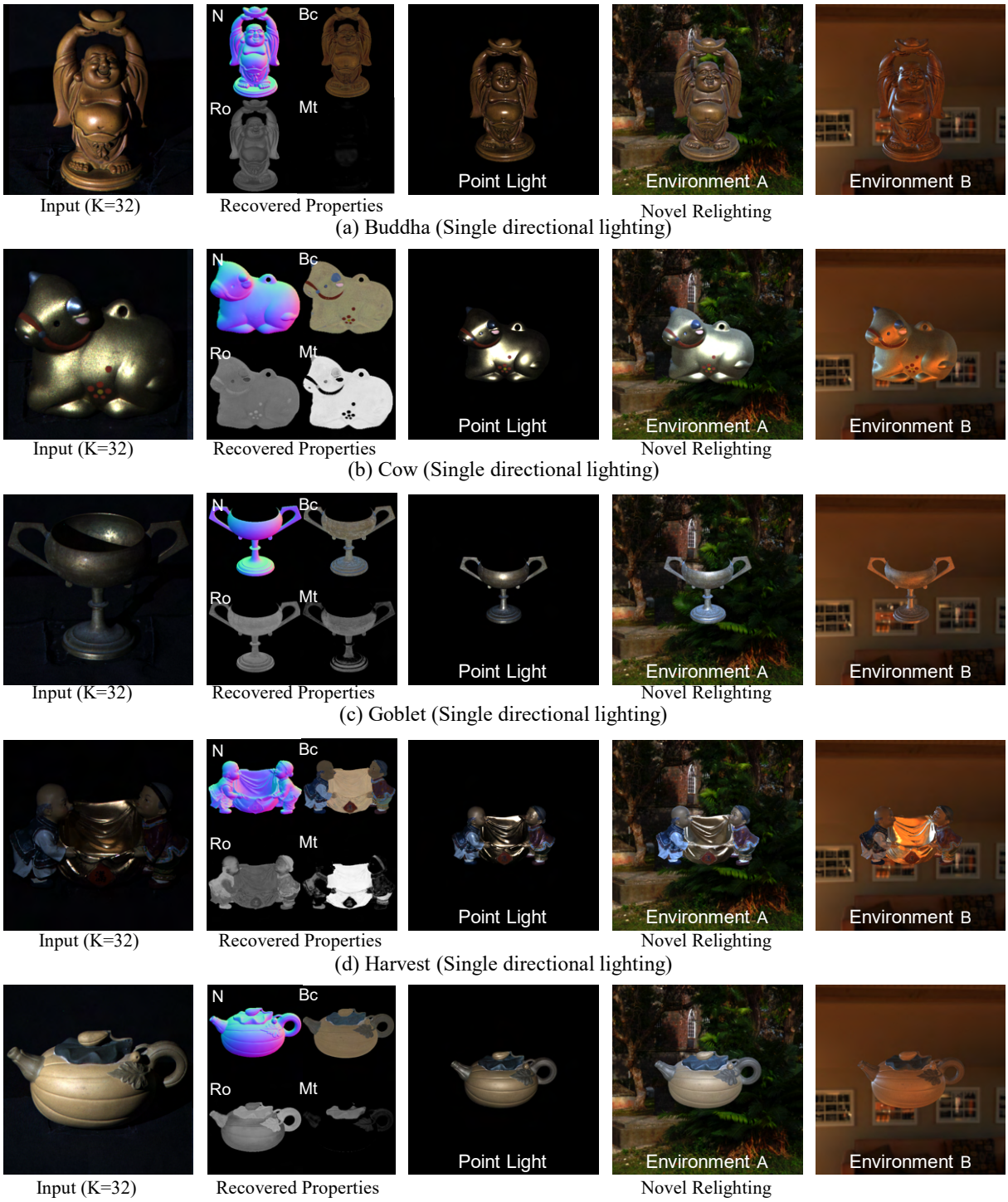
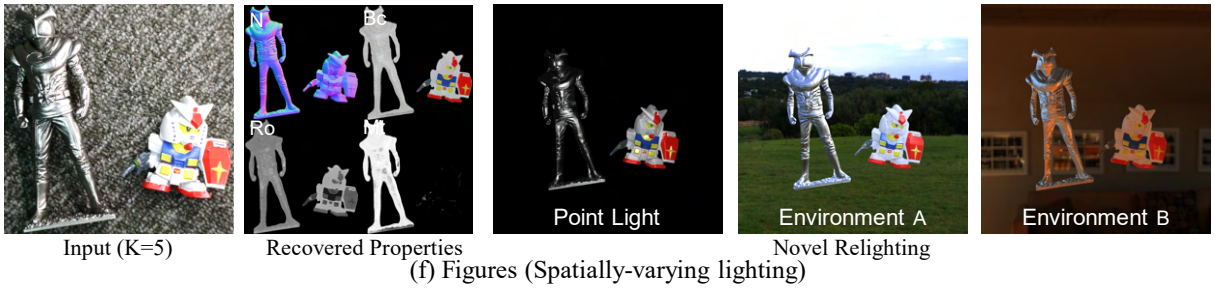


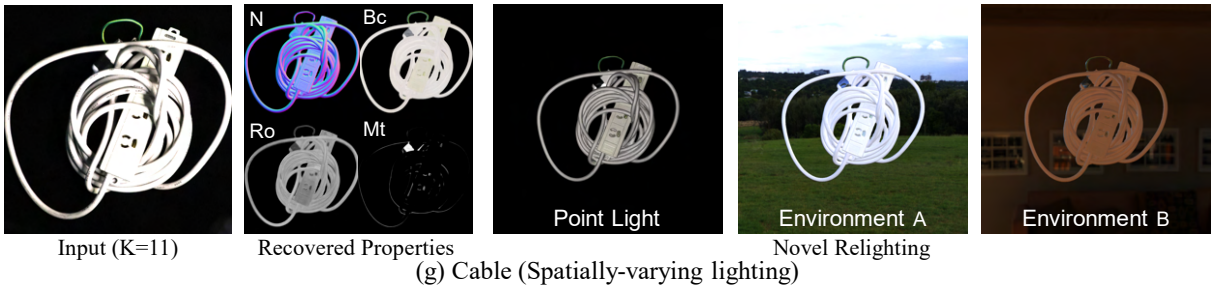
Figure 3. Reflectance recovery and novel relighting of scenes under directional lightings.

Wild [8] and our PS-Mix were rendered using the dichromatic Bidirectional Reflectance Distribution Function (BRDF) [4], which is commonly assumed in physically-based rendering of materials. This BRDF is a combination of the diffuse, specular, and metallic BRDFs, controlled by three parameters: *base color* $\in \mathbb{R}^3$, *roughness* $\in \mathbb{R}$, and *metalness* $\in \mathbb{R}$ (all parameters are within the range 0 to 1). The diffuse BRDF includes a Schlick Fresnel



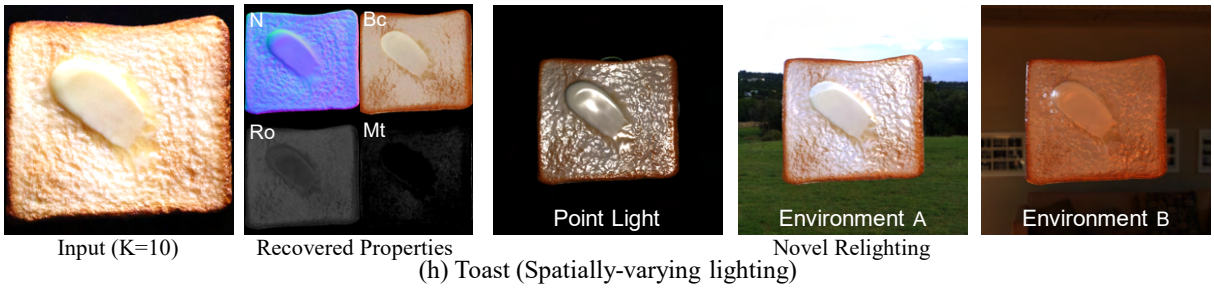
Input (K=5) Recovered Properties Novel Relighting

(f) Figures (Spatially-varying lighting)



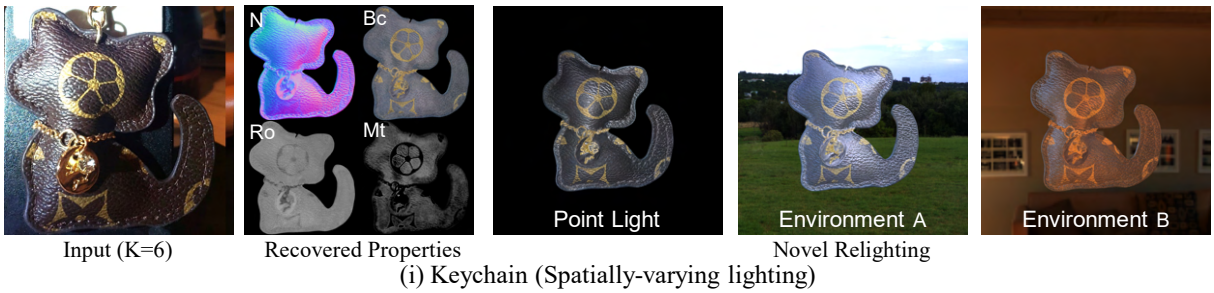
Input (K=11) Recovered Properties Novel Relighting

(g) Cable (Spatially-varying lighting)



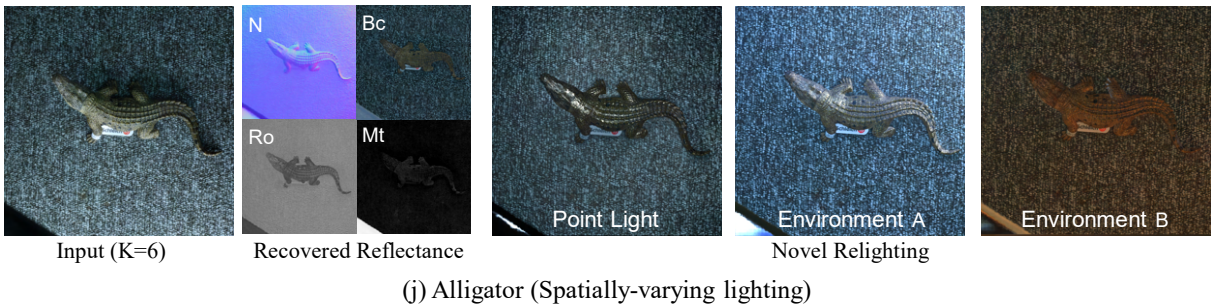
Input (K=10) Recovered Properties Novel Relighting

(h) Toast (Spatially-varying lighting)



Input (K=6) Recovered Properties Novel Relighting

(i) Keychain (Spatially-varying lighting)

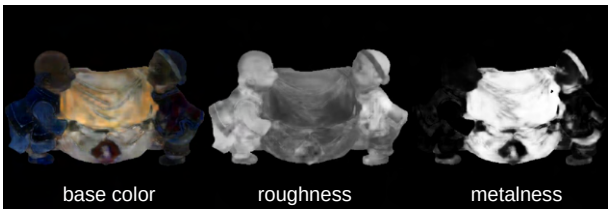


Input (K=6) Recovered Reflectance Novel Relighting

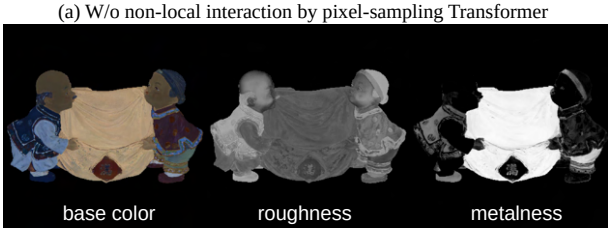
(j) Alligator (Spatially-varying lighting)

Figure 4. Reflectance recovery and novel relighting of scenes under spatially-varying illuminations.

factor and a term for diffuse retro-reflection whose color is determined by the *base color* parameter. The specular BRDF is the Cook-Torrance microfacet BRDF that uses the isotropic GGX (also known as Generalized-Trowbridge-Reitz-2) with a Smith masking-shadowing function. The *roughness* parameter controls the shape of the lobe, with smaller *roughness* values producing steeper specular lobes, i.e., more prominent specular highlights. The metallic



(a) W/o non-local interaction by pixel-sampling Transformer



(b) W/ non-local interaction by pixel-sampling Transformer

Figure 5. We compared the results of reflectance recovery w/ and w/o our pixel sampling Transformer. As we observe, the non-local interaction among aggregated features seem to be critical in the surface reflectance recovery.

BRDF uses the same specular BRDF, but the reflected light is colored with the *base color* parameter. The *metalness* parameter balances the weight between the dielectric (diffuse+specular) and metallic BRDFs. As per this definition, the metalness of a surface is primarily determined by the color of its specular reflection. In other words, if the predicted surface base color and the color of the specular reflection are similar, the surface is classified as metallic and assigned a corresponding metalness value. This can result in some black surfaces being classified as metallic, but it does not pose an issue in novel relighting since black surfaces are always represented by the same BRDF, regardless of their metalness value. For further details, please refer to [4].

Results under directional lighting: In Fig. 3, we present the results of reflectance recovery from random 32 images of five objects in DiLiGenT [15]. For each object, we show one of the input images and the recovered surface normal (N), *base color* (Bc), *roughness* (Ro), and *metalness* (Mt) maps. We observe that the proposed method could cluster identical materials, even though we did not impose any physically-based constraints on reflectance properties based on prior knowledge, such as smoothness or sparsity of basis materials, which has been done in most existing works [10, 11, 19, 20]. We also observe successful separation between the surface color and shading effects. Furthermore, several objects in DiLiGenT have metallic paintings (*e.g.*, Harvest and Cow), and our method correctly recovered the metalness values for these areas. To the best of our knowledge, our method is the first to recover physically plausible metalness parameters of non-convex scenes under unknown lighting conditions.

Using the recovered BRDF parameters, we rendered the scenes under three different lighting conditions using the physically-based renderer [2]: a point light collocated with the camera position, outdoor environment lighting, and indoor environment lighting. While the unavoidable ambiguity of the problem setup makes quantitative evaluation impossible, we obtained highly plausible rendering results for each lighting condition. Note that all results were based on the surface normal map, not the surface meshes, so we cannot render global lighting effects such as cast shadows and inter-reflections.

Our analysis of the results revealed an interesting observation: non-local interactions are more critical in recovering surface reflectance than surface normal. As shown in Fig. 5, we found that the recovery of material properties required a broad range of observations, including from low-frequency (diffuse) to high-frequency (specular) components. Reliable low-frequency information is almost sufficient for surface normal prediction, but it is not enough for recovering material properties, and focusing on a specific pixel is not adequate. The non-local interaction of aggregated features proved helpful in seeing different surface points of the same material for the recovery of surface reflectance, resulting in our outstanding results.

Results under spatially-varying lighting: In Fig. 4, we present the results of reflectance recovery and novel relighting from images under spatially-varying illuminations. The results demonstrate that the proposed method achieves physically plausible performance, overcoming the challenging conditions of each scene.

For instance, in the *Figures* dataset, a metallic-painted, non-planar object and a non-metallic, planar object exist in the same scene, but the network successfully reconstructed the normal map without distinguishing between these objects. Additionally, the *metalness* parameters were successfully recovered in the metallic-painted area, as demonstrated in the relighting results. The *Cable* objects have complex tangles of long, thin cables, and such geometries tend to produce ambiguous depth discontinuities when handled by existing methods. The proposed method not only accurately reconstructed these geometries but also recovered uniform *base color* without being affected by cast shadows or inter-reflections caused by non-convex geometries. For other objects such as *Toast*, *Keychain*, and *Alligator*, the proposed method successfully recovered a detailed surface normal map that preserved depth discontinuities accurately and produced perceptually plausible surface reflectance maps that were unaffected by shading and global illumination effects. The novel relighting results demonstrate that our results are of practical quality for the capture of surface attributes.

These results suggest that the proposed method is highly effective not only in surface normal recovery but also in surface reflectance recovery and novel relighting of the scene.

However, we emphasize once again that we are aware that the results of this experiment are not objective, and further quantitative evaluation of surface reflectance recovery is left for future work.

References

[1] Adobe Stock. https://stock.adobe.com/. 3

[2] Blender. https://www.blender.org/. 3, 6

[3] MMSegmentation. https://github.com/open-mmlab/mmsegmentation. 3

[4] B. Burley. Physically-based shading at disney, part of practical physically based shading in film and game production. *SIGGRAPH 2012 Course Notes*, 2012. 4, 6

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[7] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3

[8] S. Ikehata. Universal photometric stereo network using global lighting contexts. In *CVPR*, 2022. 3, 4

[9] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, pages 3744–3753, 2019. 3

[10] Junxuan Li and Hongdong Li. Neural reflectance for shape recovery with shadow handling. In *CVPR*, 2022. 6

[11] Junxuan Li and Hongdong Li. Self-calibrating photometric stereo by neural inverse rendering. In *ECCV*, 2022. 6

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[13] Zhuang Liu, Hanzi Mao, and Christoph Feichtenhofer Chao-Yuan Wu. A convnet for the 2020s. In *CVPR*, 2022. 2

[14] Markus N. Rabe and Charles Staats. Self-attention does not need $o(n^2)$ memory, 2021. 3

[15] B. Shi, Z. Wu, Z. Mo, D.Duan, S-K. Yeung, and P. Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *CVPR*, 2016. 6

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[17] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018. 3

[18] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR, 2020. 3

[19] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *CVPR*, 2022. 6

[20] Yuanqing Zhang, Jiaming Sun, Xinqyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. *CVPR*, 2022. 6