Supplementary Material: Learning Indoor Inverse Rendering with 3D Spatially-Varying Lighting

Zian Wang[1,2,3] Jonah Philion[1,2,3] Sanja Fidler[1,2,3] Jan Kautz[1]
NVIDIA[1] University of Toronto[2] Vector Institute[3]
{zianw, jphilion, sfidler, jkautz}@nvidia.com

In the supplementary material, we provide implementation details of our approach and analysis on additional results. We refer to the accompanied video for qualitative results of virtual object insertion on the InteriorNet test set and real world images.

1. Additional Experiment Details

Network Architecture Details. In the Direct Prediction Module, the 2D CNN backbone is ResNet18 with four branches for albedo, normal, depth and global lighting features respectively. The four branches share the first three residual convolution blocks. The downsampling module in the second and the fourth convolution block for albedo, normal and depth branches are removed, while the lighting branch is downsampled twice and passed to a global pooling layer to produce the final global lighting encoding.

For the Lighting Joint Prediction Module, to convert the global lighting feature vector into a feature volume, we follow the architecture of OccNet [7], and use three MLP residual blocks with a conditional batch normalization (CBN) [2] layer. The scene lighting feature $\mathbf{f}_L$ is decoded into a 32^3 feature volume and fused into the 3D UNet after 2 downsample convolution blocks. The 3D UNet used in the Lighting Joint Prediction Module has five downsample and upsample convolution blocks with residual connections [3], where each convolution block contains two 3D convolution layer. For the Global Feature Decoder in the Lighting Joint Prediction Module, a sequence of 3D transpose convolution can achieve similar functionality. We also tested the choice of transpose convolution and it empirically showed similar performance. We thus adopt the MLP module, which is more flexible and can easily extend to multi-view input.

For the Joint Re-prediction Module, we use a 6-conv-block 2D UNet [6] for reflectance and shape joint re-prediction. For the UNet architecture, each convolution block in the first half downsamples 2x spatial resolution while the rest upsample 2x, connected by residual connections [3].

Training Details. We set λ_{adv}, λ_{reg} to 3e-3 and 1e-3, and other loss ratios are set to 1. We train each module with Adam [4] for 100 epochs each, with learning rate as 3e-4 decaying by 0.3 every 30 epochs, and then we jointly finetune for 30 epochs with learning rate as 1e-5.

Virtual Object Insertion provides an important evaluation of lighting estimation. Our editing process involves two parts: (1) rendering the appearance of the foreground inserted object, and (2) rendering the residual effects for background scene image I_{scene} due to the inserted object, such as cast shadows.

For rendering the appearance of inserted objects, we directly do raytracing and query the radiance of each ray from the predicted lighting representation. To edit image pixels, we measure the residual effects caused by the inserted object and apply it to the background scene pixels. Specifically, we compute Lambertian shading before and after the insertion of the virtual object, denoted as S_{before} and S_{after}. We use a ratio $r = \frac{S_{\text{after}}}{S_{\text{before}}}$ to represent the effect caused by the inserted object. We reshade the scene image by $I_{\text{edit}} = rI_{\text{scene}}$. Geometry of the visible surface comes from our predicted depth, and shading is computed based on the predicted normals and 3D spatially-varying lighting. For occluded rays when computing S_{after}, we perform one bounce ray-tracing to get the new radiance.

For the purely specular objects in our paper, we use Phong glossiness factor as 512. For the mostly diffuse objects, we set RGB diffuse albedo as 0.9, and use 0.1 specular reflection with Phong glossiness as 32.

When comparing with Li *et al*. [5] which only estimates 2D spatially-varying lighting, we select the nearest 2D surface point and use the lighting prediction at the 2D surface point.

Experiment Settings of Funetuning on Real-world Data. When evaluating on real-world testing sets, we also finetune our model on corresponding training sets. To ensure a fair comparison, the training procedure is consistent between our method and baselines for all experiments shown in the paper.

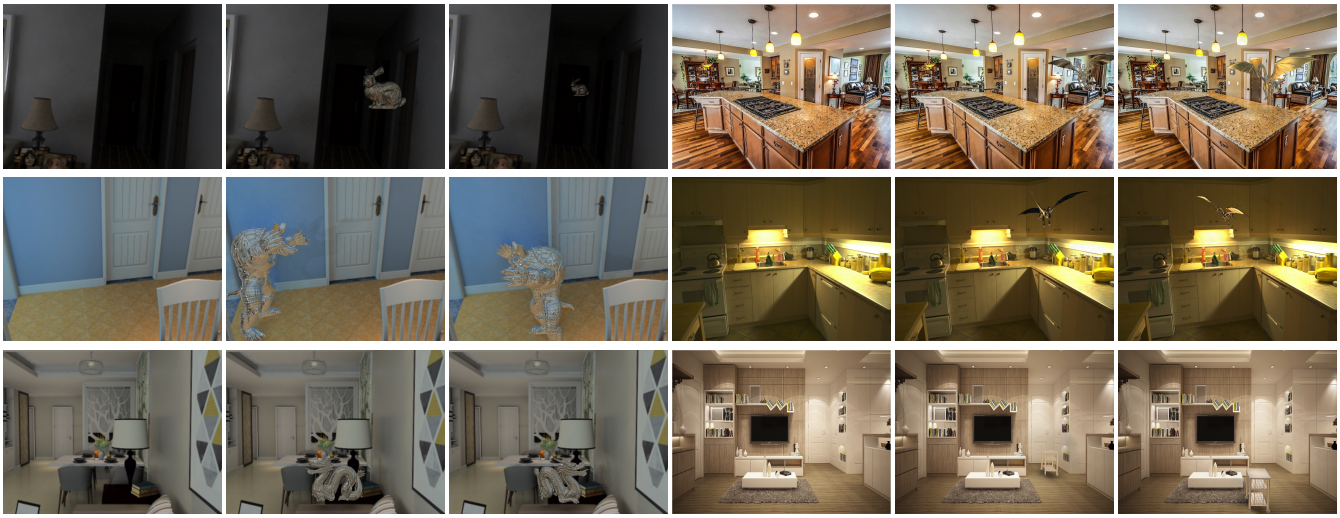

Figure A: **Qualitative Results of Virtual Object Insertion.** For each example, from left to right are input image and two edited results with the inserted virtual object at different location.

Figure B: **Qualitative Results of Transparent Object Insertion.** From left to right are input image and inserted transparent objects.

We finetuned and evaluated albedo on IIW dataset [1]. IIW dataset provides pair-wise sparse human annotation on albedo. Specifically, the sampled point pair P_1, P_2 is presented to the annotator, and the annotator gives the judgement of relative relationship of the albedo values A_1, A_2, *i.e.* A_1 is greater than, equal to or less than A_2. Following [9, 10], we use a hinge loss based on this annotation. In our experiment, we first pretrain our method and the baseline method on InteriorNet and then jointly train with both InteriorNet and IIW supervision. For normals and depth, we finetuned and evaluated on NYUv2 [8] dataset. We use the same loss for normals and depth as on InteriorNet. Similar to experiment settings on IIW, we first pre-train on InteriorNet and jointly train with NYUv2 supervision. We also collected 120 LDR real-world panoramas of indoor scenes from the Internet, and jointly train with InteriorNet. During training, the 120 LDR panoramas are randomly rotated as a data augmentation. For each transformed panorama, we use InteriorNet camera intrinsics to crop out a perspective image as network input. With the predicted lighting volume, we render the predicted panorama at camera center and enforce consistency with GT panorama using a L2 loss. We also use the real-world panoramas to train the discriminator. We show in the main paper that finetuning on LDR panoramas improves the performance of our method. With the growing interests on VR and portable panorama capturing devices, we believe the data collection of panoramic images and videos will be much easier, and will greatly benefit our method.

2. Additional Results and Analysis

Qualitative Results of Virtual Object Insertion. Fig. A shows the results of inserting challenging, purely specular objects into indoor scenes. As shown in the top-left bunny example, our predicted lighting is 3D spatially-varying and can capture lighting intensity changes within the scene. In the top-right and middle-right dragon examples, the insertion results show that our model correctly captures geometry of light sources and produces spatially consistent specularities. Our HDR output enables cast shadows of the rendered objects. In the middle-left, bottom-left and bottom-right examples, the inserted objects lead to strong cast shadows onto the wall and the floor, which is consistent with the visual cues in the input images.

As shown in Fig. B, our estimated lighting can produce realistic insertion results for transparent objects, which indicates that our predicted lighting volume preserves high frequency details.

Figure C: Appearance and geometry visualization of the predicted lighting volume.

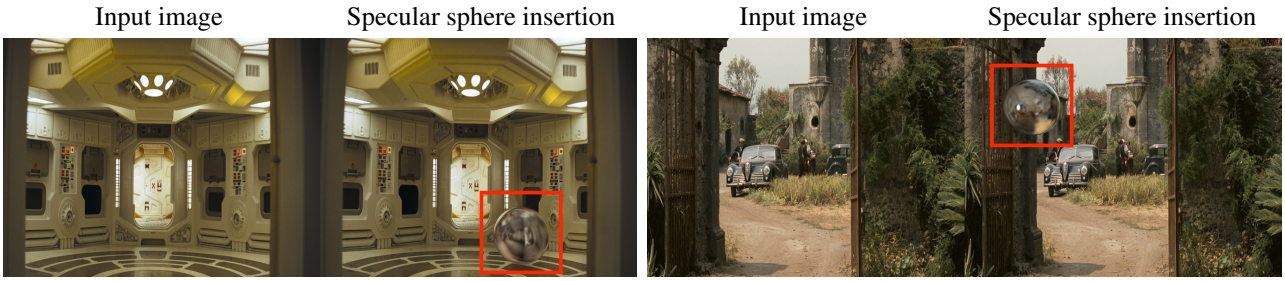

Figure D: Direct inference results on out-of-domain imagery. From left to right, we show input images and editing results of inserting a purely specular sphere.

Qualitative Comparison of Specular Sphere Insertion. We qualitatively evaluate lighting estimation on both InteriorNet and real-world images, and show the editing results in our *accompanied video*. We compare with prior works and our ablated model by inserting a highly specular sphere into the scene, which faithfully reflects the quality of lighting prediction.

Compared to NIR [10] and Li *et al*. [5], our method can capture both 3D spatially-varying lighting and angular high-frequency details, while NIR uses a single environment map that hardly captures lighting variation, and Li *et al*. severely suffers from spatial instability. Both NIR [10] and Li *et al*. [5] cannot recover angular high-frequency details. Compared to Lighthouse [11], our lighting prediction can recover more HDR information, and can produce realistic cast shadows while Lighthouse is not possible to do so. Our lighting prediction also shows more intensity variation while Lighthouse predicts almost uniform lighting intensity. Note that Lighthouse is using stereo pair as input instead of monocular image, which includes more information for the visible FoV.

We also show the qualitative results of our model trained without Re-rendering Loss. When training without the Re-rendering Loss, our model shows similar effects as Lighthouse [11], which produces less lighting variation and cannot predict HDR lighting. This further validates the effectiveness of our holistic inverse rendering framework and the re-rendering loss for joint reasoning.

We also compare with NIR [10] and Li *et al*. [5] on real-world images. We did not compare with Lighthouse as it requires a stereo pair as input. The results further demonstrate that our method generalizes well to real-world images and outperforms prior methods.

Visualization of predicted lighting volume. We visualize the appearance and geometry prediction of the lighting volume in Fig. C. The predicted lighting volume preserves appearance and geometry of visible FoV, with only minor loss due to voxel resolution. The rendered panoramas show reasonable appearance and geometry to generate realistic results for downstream tasks. Interestingly, as shown by the rendered depth panorama, the predicted outside-FoV geometry is also reasonable. Despite only RGB supervision for outside FoV, the outside-FoV geometry prediction potentially benefits from the translation equivariance of 3D convolution, where the learned geometry prior for the visible FoV also regularizes the outside-FoV prediction.

Limitation and future work. Similar to other deep learning models, when directly applying our model to scenes that are significantly out-of-domain, the model might get results biased by the training set. As shown in Fig. D, the predicted lighting may still mimic the training data and predict indoor high-frequency appearance. Interestingly, our model still correctly predicts the highlight direction (Fig. D right), which indicates benefits of our physics-based design. Another limitation of our method is the Lambertian surface assumption. It is an interesting future work to model complex materials such as specular and transparent surfaces, which is crucial for complex visual effects in applications such as scene relighting.

References

[1] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):159, 2014. 2

[2] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2017. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 1

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 1

[5] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020. 1, 3

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[7] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 1

[8] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[9] Thomas Nestmeyer and Peter V Gehler. Reflectance adaptive filtering improves intrinsic image estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6789–6798, 2017. 2

[10] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[11] Pratul P Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8080–8089, 2020. 3