Channel-wise Noise Scheduled Diffusion for Inverse Rendering in Indoor Scenes

Supplementary Material

In this supplementary material, we first demonstrate the complementary nature of PDM and SDM, highlighting the importance of considering both accuracy and diversity in inverse rendering (Appendix A). Second, we provide the implementation details (Appendix B). Next, we present the application results (Appendix C), and finally, we provide additional experimental results (Appendix D).

A. Validation of the Complementarity

As shown in Figs. 7, 8, 9, and 10, our PDM exhibits high variance in ambiguous regions, making it a useful measure for assessing uncertainty in inverse rendering. We leveraged this characteristic to validate the complementarity between PDM and SDM. In Fig. 1, we measured the variance and error of PDM across datasets and conducted linear regression to analyze the correlation between error and variance. Each graph includes the Pearson correlation coefficient in the lower right corner. As expected, a higher variance led to a higher error (lower accuracy) across all datasets. This is confirmed by the positive Pearson coefficients, indicating that PDM presents diverse solutions for ambiguous regions. Furthermore, SDM (blue line) excels in low-variance images, while PDM (red line) outperforms in ambiguous, high-variance images. This fact demonstrates the complementary nature of PDM and SDM, emphasizing the importance of incorporating both accuracy and diversity in inverse rendering. For the roughness map of OpenRooms FF, due to the inherent uncertainty of the modality, PDM consistently outperformed SDM.

For a more thorough validation, in Tab. 1, we measured the average variance of N, D, A, and R for each dataset. OpenRooms FF, where SDM outperforms (see Tab. 3 (main paper)), exhibits low variance, indicating that it contains a large number of straightforward images. In contrast, the high variance of MAW and DIODE reflects ambiguity in complex real-world images. As shown in Tabs. 5 and 6 (main paper), PDM proves to be a useful choice for handling such ambiguous images.

Dataset	OpenRooms FF [1]	MAW [11]	Diode [9]
variance	0.039	0.179	0.184

Table 1. **PDM sample variance ($\times 10^{-2}$) for each dataset.**

B. Implementation Details

All experiments were performed on eight A5000 GPUs and AdamW [7] optimizer was used for all models. Next, we provide the implementation details for each model. Each model was trained separately and subsequently frozen.

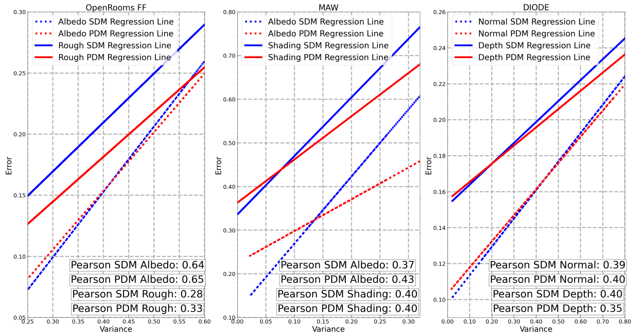

Figure 1. **Correlation of variance(x-axis) and error(y-axis).** The error of PDM was calculated from the mean predictions. The blue line represents SDM, while the red line represents PDM.

B.1. Implicit Lighting Representation

The input per-pixel environment map E is flattened on a pixel basis, followed by the application of `log1p`. The encoder consists of 7 layers, each with 64 hidden units, followed by a 1D BatchNorm. The feature vector $\boldsymbol{f}$ is constrained to the range $[-1, 1]$ by a `tanh` activation and has a dimensionality of 96. The decoders Decoder_E, Decoder_S, and Decoder_I are each composed of 3, 3, and 6 layers with 128 hidden units, respectively, without any normalization layers. Furthermore, to improve the robustness of Decoder_S, random roughness values were generated and used during training. The loss function consists of a log-space L_2 loss for E, S, I_S, I_S (random), and a standard L_2 loss for I. The model was trained with a batch size of 256 for 30 epochs, which took approximately one day to complete. These encoders and decoders are jointly trained and are subsequently frozen.

B.2. Diffusion-based Inverse Rendering

The DM was based on the `UNet2DConditionModel` from Diffusers [10], which consists of 586M parameters. We applied classifier-free guidance [3] with a 0.05 probability and rescaled the noise schedule following Lin *et al.* [6]. The training was carried out with a batch size of 512 for 150 epochs, which took approximately 2 days to complete. After training, PDM used 10 DDIM [8] inference steps.

B.3. RGB-Guided Super Resolution

The SRM's encoder includes an encoder identical to the DenseNet [4] structure used for I, which produces dense feature maps. In addition, a separate encoder for $\mathbf{z}_0$ is provided. This encoder consists of a single DenseBlock, a component of DenseNet, and its output is later combined with the dense feature maps as input to the modality-specific decoders. Each decoder also receives the corresponding low-resolution modality as input. The decoders follow the same structure as the existing baselines [5]. The SRM was trained with a batch size of 128 for 50 epochs, taking approximately 2 days to complete. The loss function consists of an L_2 loss for N, a scale-invariant log-space L_2 loss for D, a scale-invariant L_2 loss for A, an L_2 loss for R, and an L_1 loss for $\boldsymbol{f}$.

Performance Analysis of SRM. For PDM, there was a tendency for slight residual noise to remain even after training. We analyzed whether SRM effectively removes this noise while preserving the identity of the samples generated by PDM. Fig. 2 presents the results of this analysis. Both the top and bottom low-resolution samples in Fig. 2 contain slight residual noise. However, in both cases, SRM successfully removes this noise and performs upsampling while maintaining the identity of each modality.

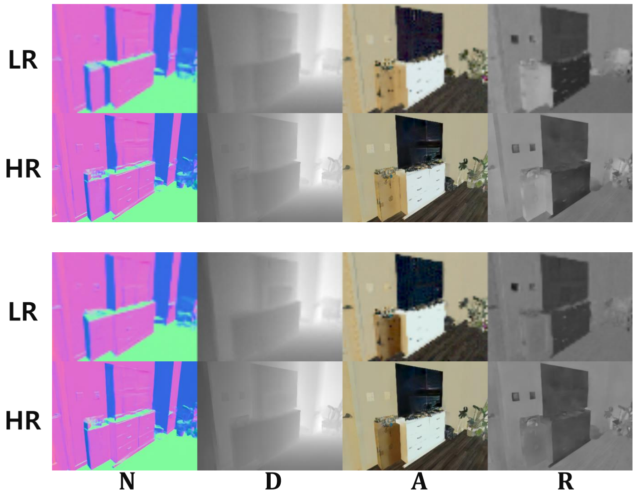

Figure 2. Performance Analysis of SRM. LR refers to the low-resolution sample, while HR denotes the high-resolution sample processed by SRM.

C. Applications

C.1. Material Editing

Inverse rendering, which decomposes an image into geometry, material, and lighting, enables applications such as material editing. In particular, ILR-based neural rendering allows for realistic rendering without the need for an external renderer. Fig. 3 demonstrates this capability. In each image, we modified the albedo to green, light purple, and pink, respectively, and for the third image, we reduced the roughness to introduce additional specularity. In all results, the spatially-varying lighting of the input image is faithfully reproduced, and the materials are realistically modified.

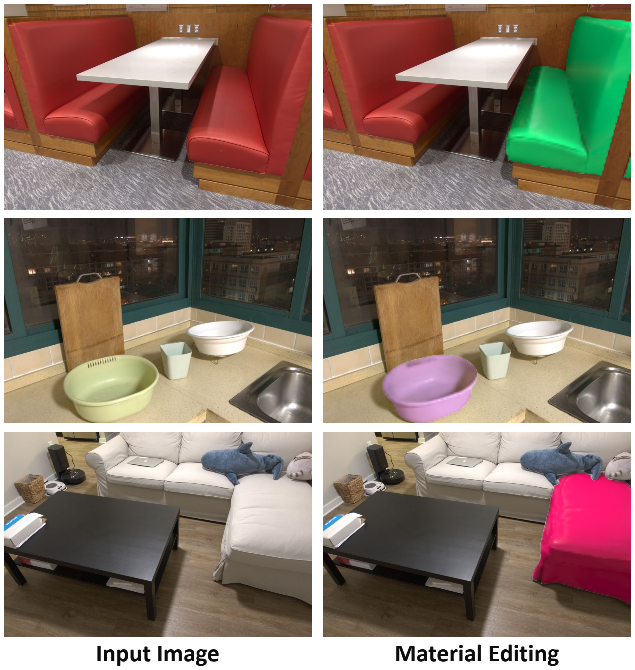

Figure 3. Material editing results.

C.2. Object Insertion

Figs. 4, 5, and 6 present the object insertion results for all 20 images from the spatially-varying lighting dataset [2]. In the user study, participants were asked to evaluate four object insertion images and respond to the question: ”Of the four images, please select the one that shows white rabbits blending naturally into the surrounding light.” The numbers in the top-left corner of each image indicate the selection rate from the user study, which was conducted with 194 participants. In most cases, our method was preferred by more users compared to baselines. Interestingly, for some images, our results were even preferred over the ground truth. We attribute this to potential discrepancies between the lighting visible in the limited field-of-view RGB image and the actual lighting in the real scene.

D. Additional Experimental Results

We have included additional experimental results for each dataset.

D.1. Synthetic Scenes

Fig. 7 presents additional experimental results in the synthetic dataset. In the upper sample, strong shadows are cast on the floor, which the baselines fail to remove. In contrast, our SDM and PDM successfully eliminate these shadows.

Similarly, for the table in the lower sample, which exhibits strong specular radiance, our method accurately extracts the diffuse albedo.

D.2. Real-World Scenes

Fig. 8 shows the inverse rendering results on images from the real-world spatially-varying lighting dataset. In the first sample, Li *et al.* [5] overestimated the albedo of the chair and wall, making them appear too bright, while Zhu *et al.* [12] introduced blotchy artifacts. In contrast, our method accurately predicted the dark regions of the wallpaper and the black albedo of the chair. In the second sample, our method cleanly predicted the albedo of the white mat on the floor. Li *et al.* [5] incorrectly predicted it as black and Zhu *et al.* [12]displayed artifacts in the albedo of the white wall. In the third sample, Li *et al.* [5] failed to predict the geometry of the floor and Zhu *et al.* [12]introduced artifacts, while our method consistently predicted both the geometry and albedo. For the fourth sample with strong directional lighting on the white desk, our method predicted a spatially consistent albedo, whereas the baselines exhibited artifacts.

D.3. Intrinsic Decomposition

Fig. 9 includes additional experimental results on intrinsic decomposition. For the brown carpet in the bottom-right corner of the first sample, unlike the baselines, which predicted an overly bright albedo, our method accurately predicted it. In the second sample, the baselines failed to remove the specular radiance from the brown table, whereas our method effectively removed it. In the third sample, Li *et al.* [5] did not remove the specular radiance from the central table and Zhu *et al.* [12]showed artifacts in the albedo of the black sofa. In the fourth sample, the baselines struggled to remove shadows from the wall, while our method predicted a spatially consistent albedo.

D.4. Geometry Prediction

Fig. 10 presents additional experimental results on geometry prediction. For the cabinet in the first sample, unlike the baselines, which exhibited significant artifacts, our method predicted the geometry relatively accurately. In the second sample, the baselines struggled to understand the context and produced inconsistent geometry predictions. In the third sample, our method demonstrated the fewest artifacts. In the fourth sample, while the baselines displayed numerous artifacts in the normal map predictions for the wall, our method provided comparatively accurate results.

D.5. Sample Diversity

Figs. 11 and 12 present experiments that evaluate how effectively PDM can provide a diverse set of possible solutions for inverse rendering. Through the diverse samples generated, we also confirmed that PDM accounts for the dependencies between multiple modalities.

References

[1] JunYong Choi, SeokYeong Lee, Haesol Park, Seung-Won Jung, Ig-Jae Kim, and Junghyun Cho. Mair: multi-view attention inverse rendering with 3d spatially-varying lighting estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8392–8401. IEEE, 2023. 1, 7

[2] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast spatially-varying indoor lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2019. 2, 4, 5, 6, 8

[3] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[4] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 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. 2, 3

[6] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5404–5411, 2024. 1

[7] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[9] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 1, 10

[10] Patrick von Platen, Suraj Patil, Anton Lozhkov, Anton Osokin, Kashif Rasul, Nathan Lambert, et al. Diffusers: State-of-the-art diffusion models for image and audio generation in pytorch, 2022. Version 0.12.0. 1

[11] Jiaye Wu, Sanjoy Chowdhury, Hariharmano Shanmugaraja, David Jacobs, and Soumyadip Sengupta. Measured albedo in the wild: Filling the gap in intrinsics evaluation. In *2023 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2023. 1, 9

[12] Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia 2022 Conference Papers*. ACM, 2022. 3

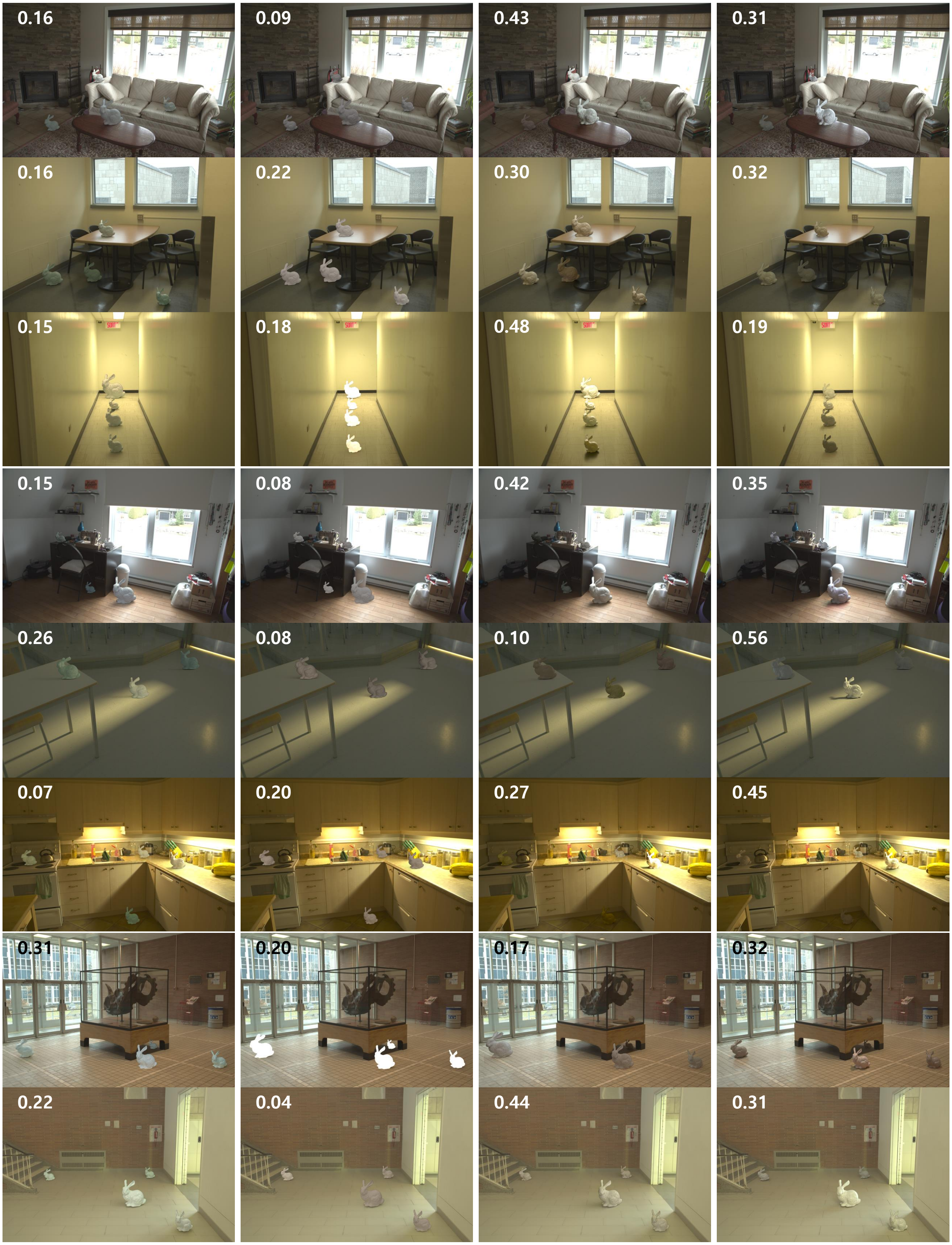

Figure 4. Additional object inversion results on spatially-varying lighting dataset. [2].

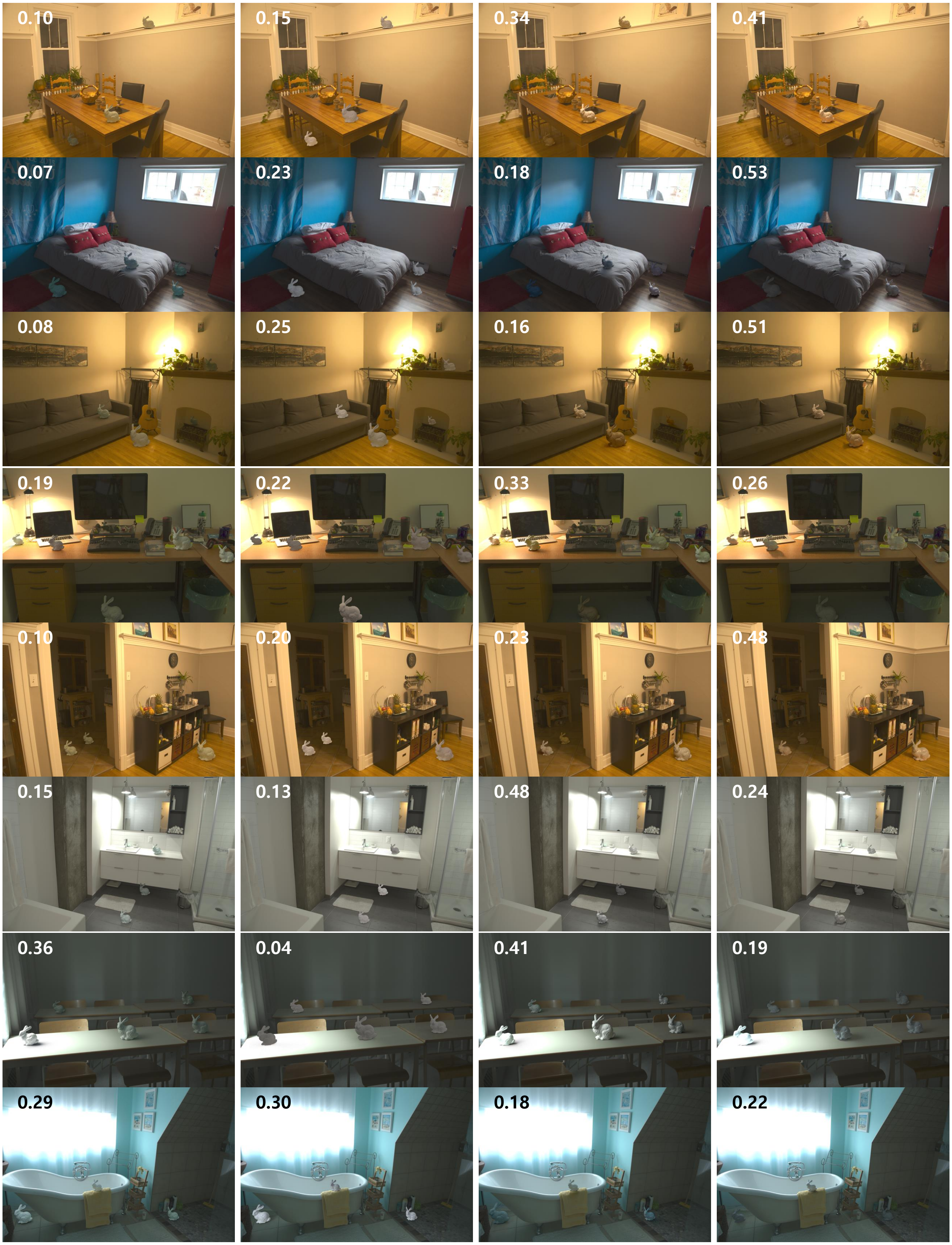

Figure 5. Additional object inversion results on spatially-varying lighting dataset. [2].

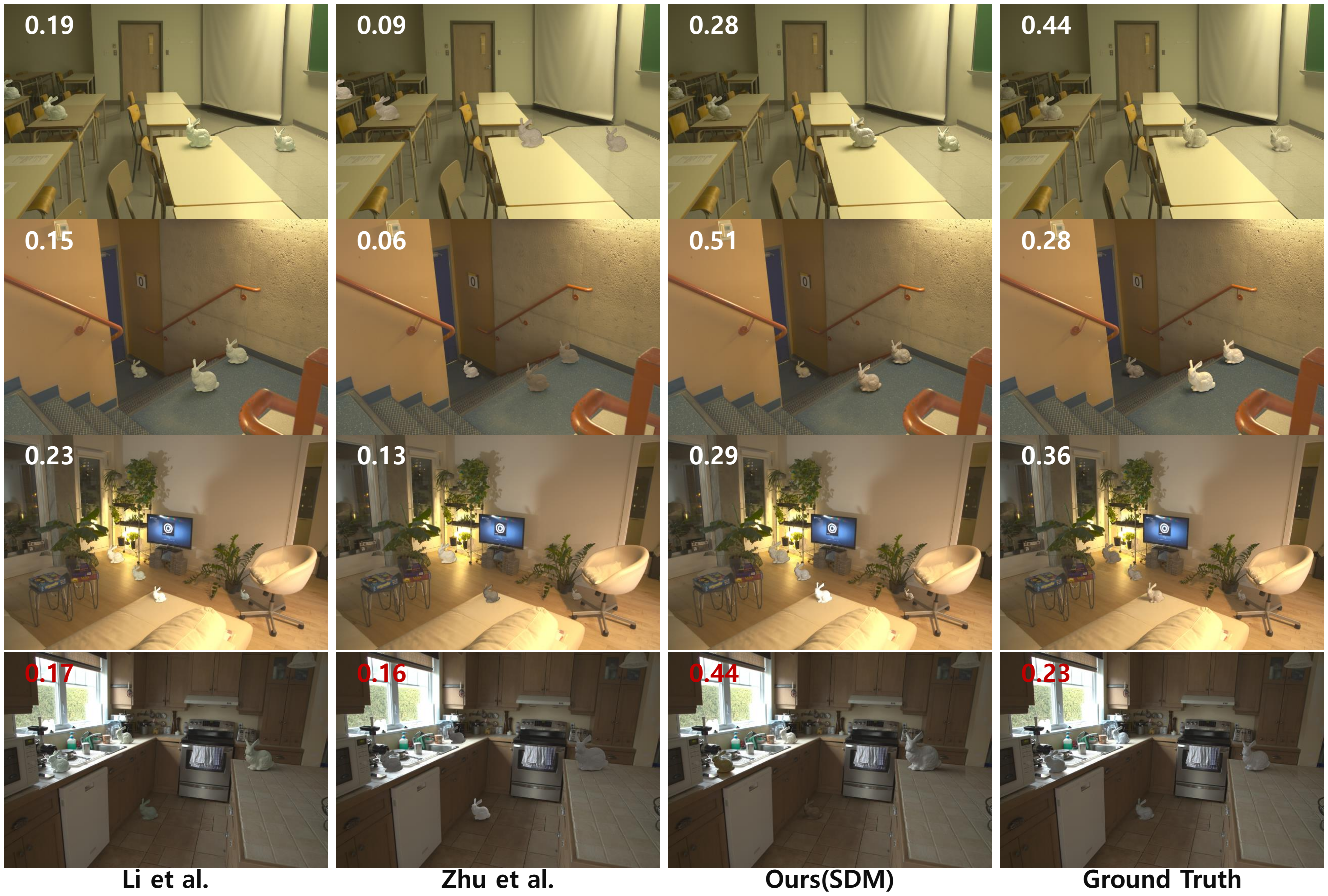

Figure 6. Additional object inversion results on spatially-varying lighting dataset. [2].

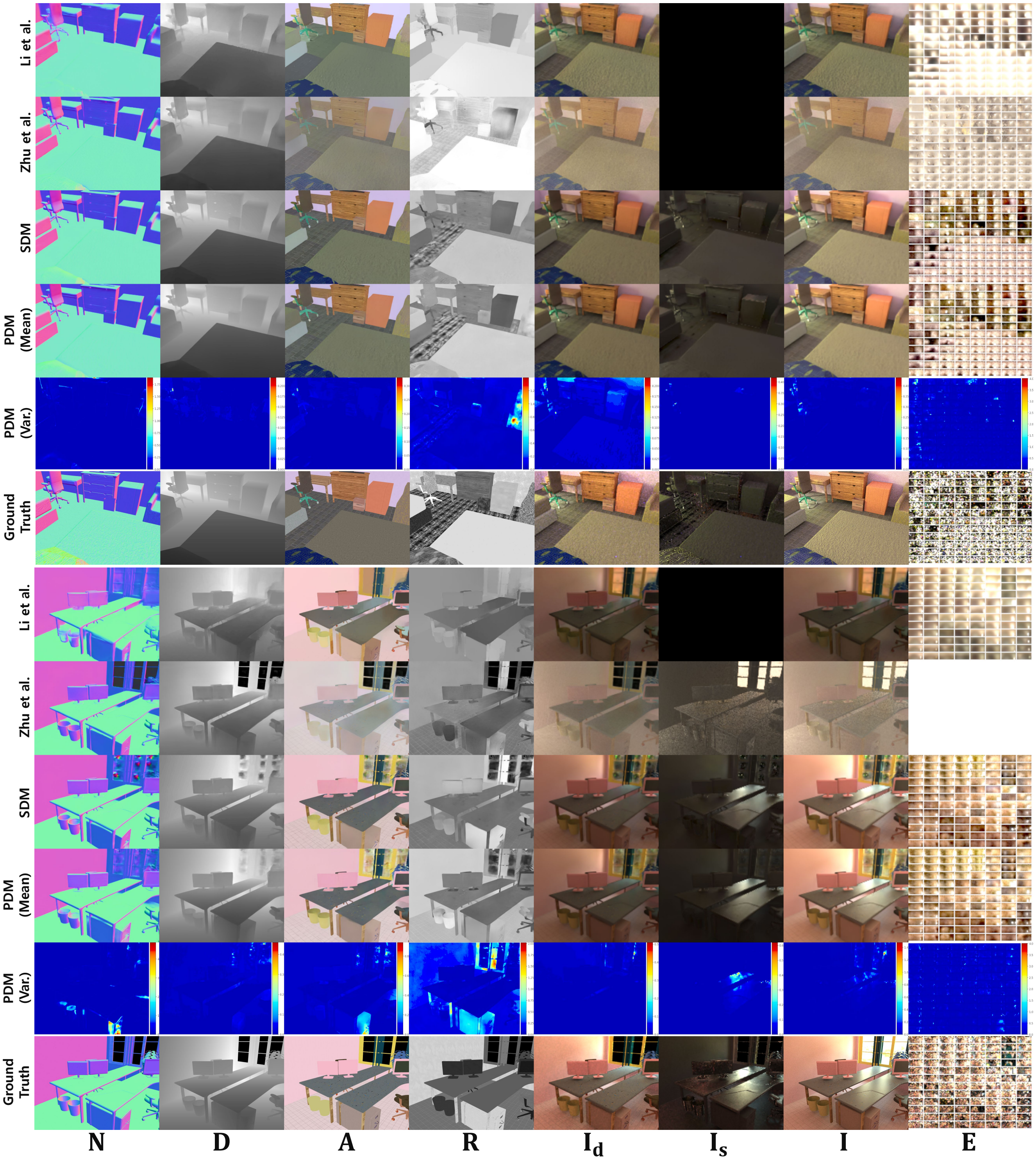

Figure 7. Additional inverse rendering results on OpenRooms FF dataset [1].

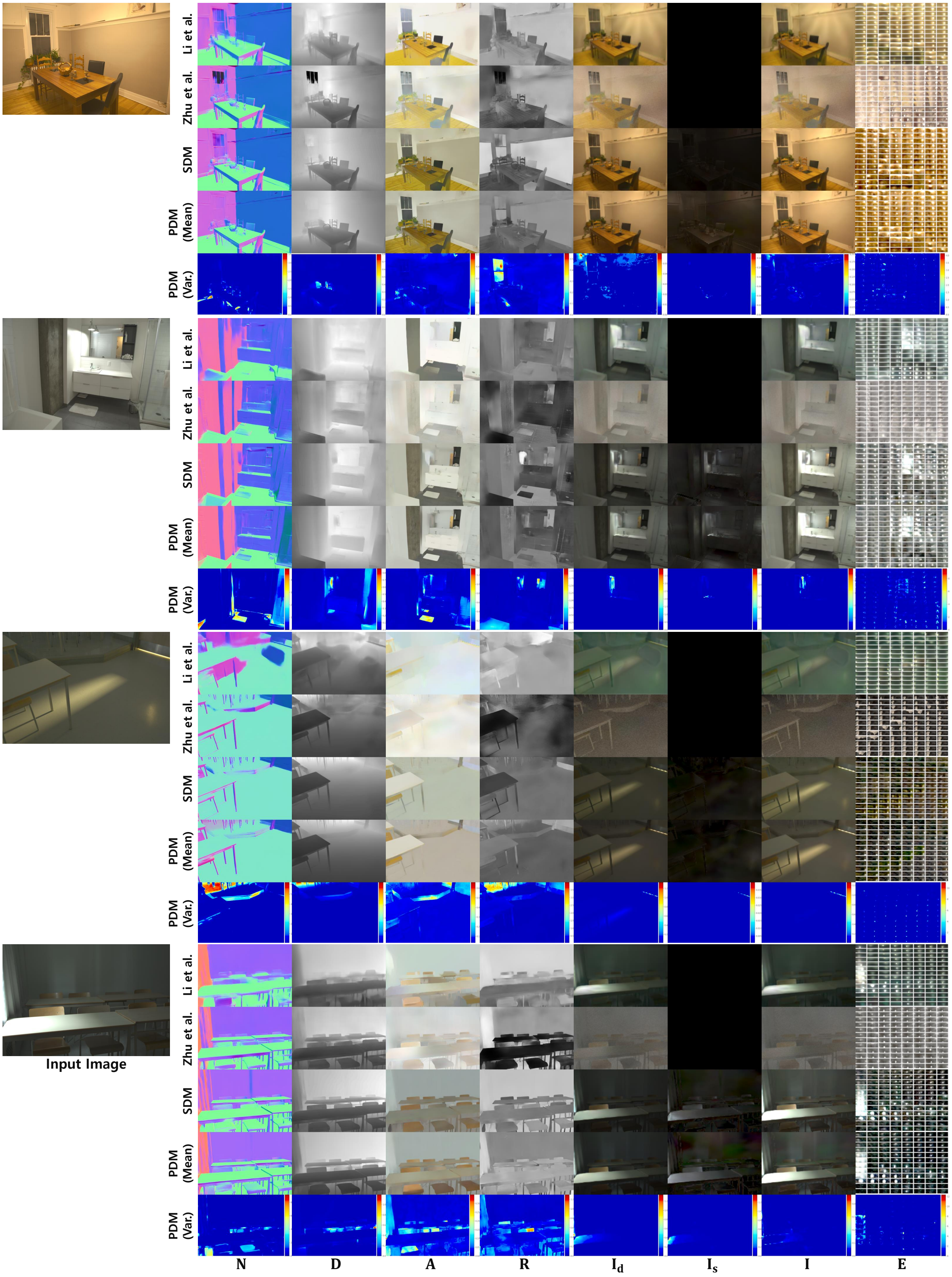

Figure 8. Additional inverse rendering results on spatially-varying lighting dataset. [2].

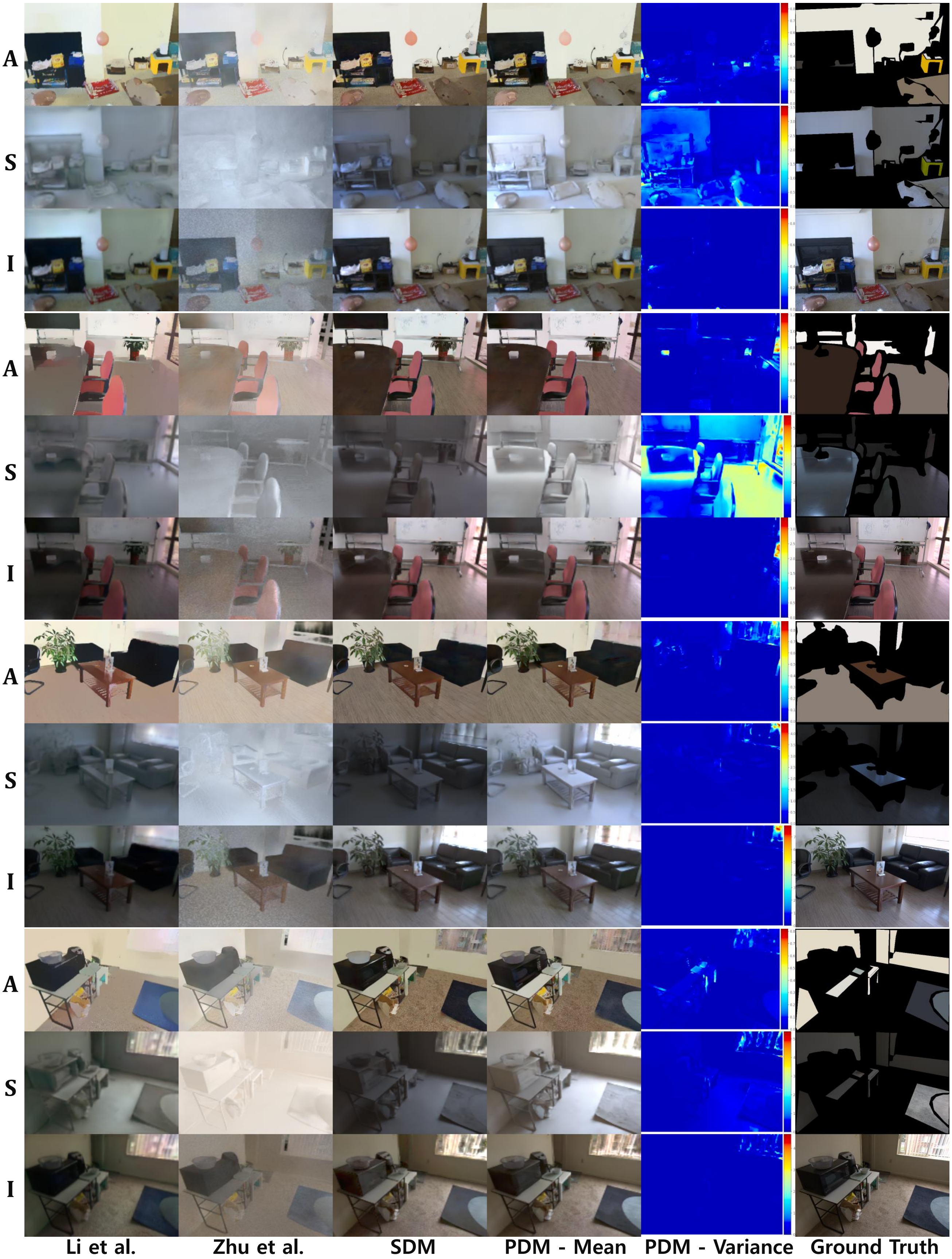

Figure 9. Additional intrinsic decomposition results on MAW dataset [11].

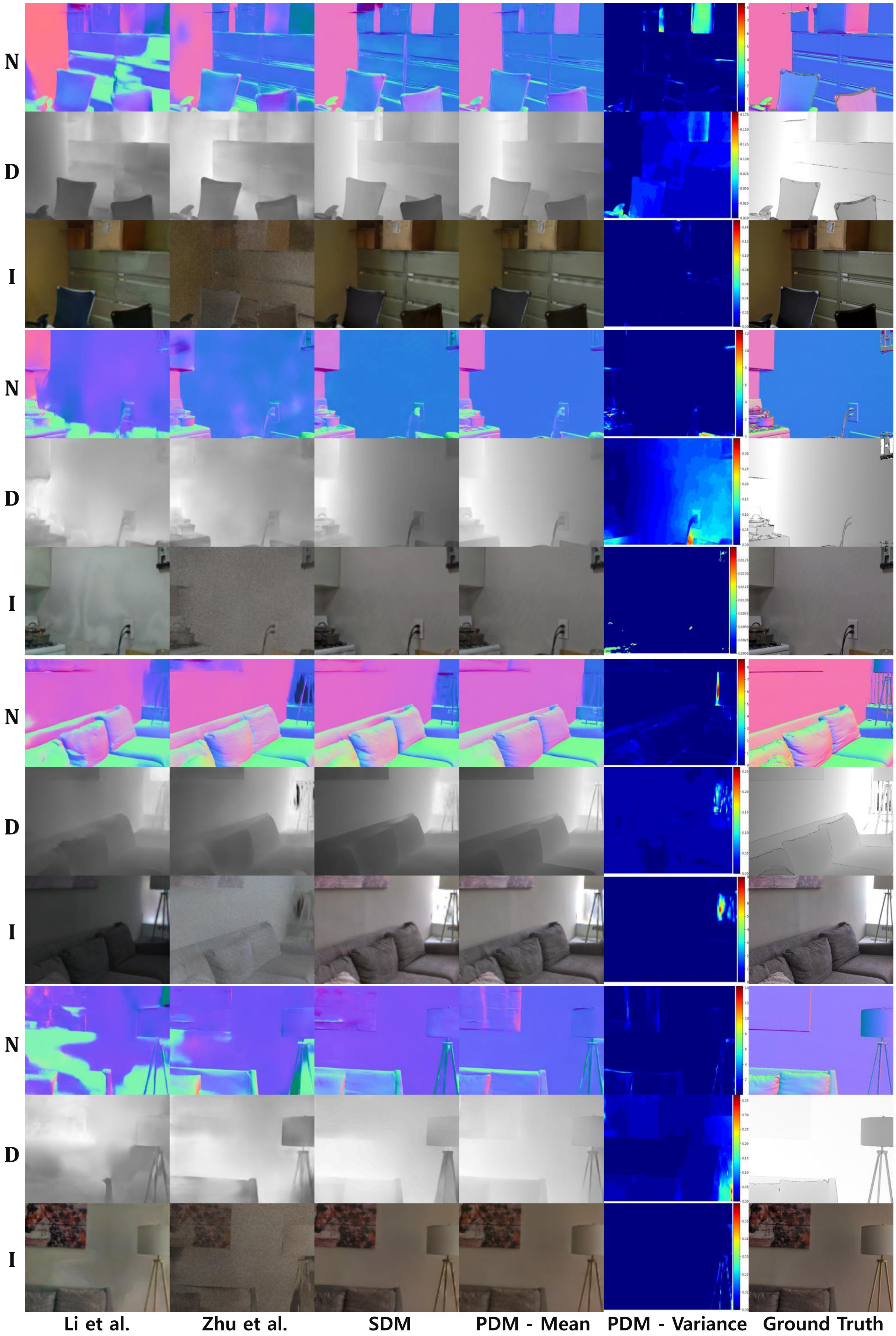

Figure 10. Additional geometry prediction results on DIODE dataset [9].

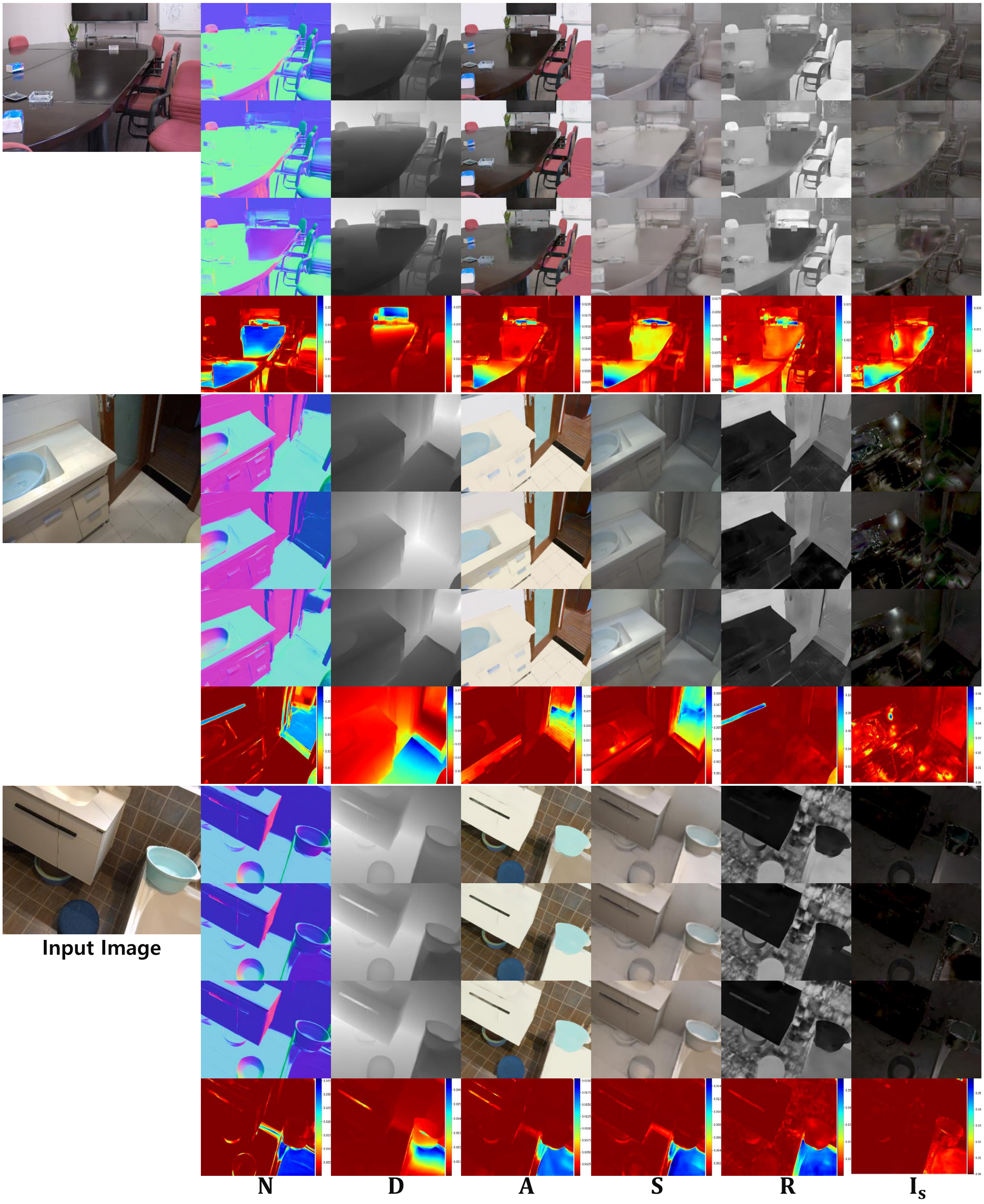

Figure 11. Additional experiments on the diversity of samples generated by PDM.

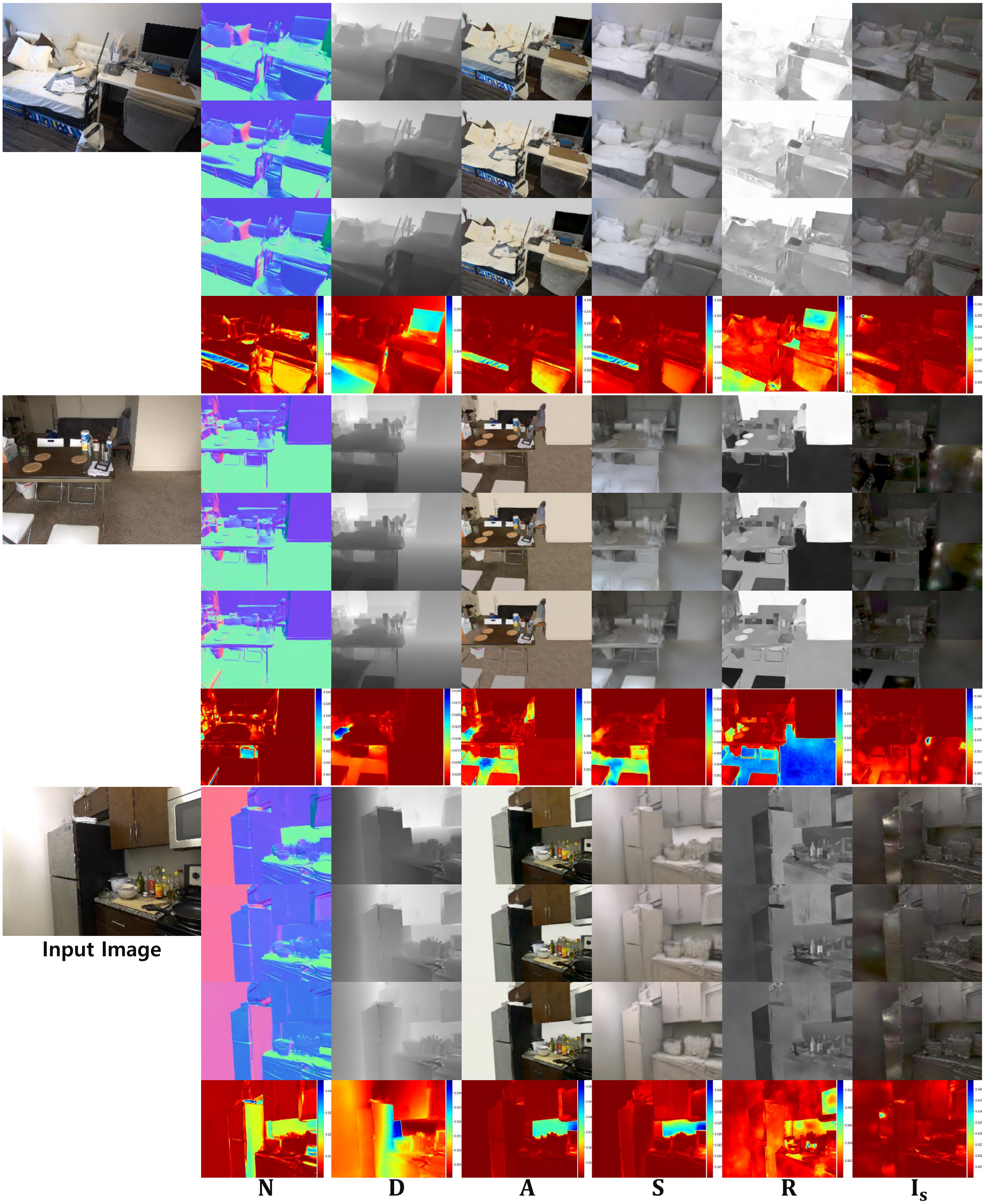

Figure 12. Additional experiments on the diversity of samples generated by PDM.