Intrinsic Image Diffusion for Indoor Single-view Material Estimation

Supplementary Material

In this supplementary material, first, we show application results (Appendix A). Then, we give additional details on our method (Appendix B) and on the experimental setting (Appendix C). We show more qualitative results on the full material prediction, including the albedo, roughness, and metallic properties on synthetic and real data in Appendix D. Finally, we present an additional ablation (Appendix E).

A. Applications

Our predicted materials and optimized lighting enable intrinsic image editing, i.e., changing solely specific aspects of the image, such as only the materials or lighting.

A.1. Material Editing

Our sharp material predictions have smooth but sharp features for every object with minimal or no baked-in lighting, enabling simple image editing in the material space. An example editing is shown in Fig. 10, where we change the wall color from beige to cyan. Note how the reflections on the wall turn greenish since the lamp emission mostly contains red and green components, but the wall reflects green and blue the most.

A.2. Lighting Editing

Our lighting provides a flexible yet controllable way to represent lighting in the scene. After fitting, the emission weights of the light sources can be edited independently. Thanks to the emissive representation, we can achieve physically realistic relighting (Fig. 10).

Figure 10. **Applications.** We show two key intrinsic image editing applications. Our method produces sharp and consistent materials without baked-in lighting, enabling convincing material editing. Our lighting representation allows fitting to small light sources, even close to objects, and editing them independently.

B. Method Details

B.1. Lighting Optimization

We use a hybrid lighting representation. For global and out-of-view lighting effects, we use a pre-integrated environment lighting parametrized by Spherical Gaussians (SG) [27]. However, such representation alone is not sufficient in our case since indoor scenes often have multiple light sources close to objects, even with different emission profiles and varying colors requiring a spatially-varying lighting representation. To achieve a controllable yet expressive representation, we additionally use N_{light} point light sources. We use SG emission profile for the point lights to further improve the expressivity.

Specifically, for the global environment map and also for each point lights, we use N_{sg} SGs with separate 3-channel weights. The point lights positions are initialized over a grid in image space and are backprojected to 3D with $1e{-}2$ offset from the surface in the normal direction in normalized depth space. The emission profiles are initialized with minimal uniform emission.

We use our predicted materials, and OmniData [11] normal estimation to rerender the scene and optimize the lighting parameters with L2 reconstruction loss. We consider every light source for each pixel but without occlusions. We found that using the absolute value of the geometry term makes the optimization more stable because otherwise if a light source happens to move behind an object out of the scene, it receives no gradients anymore.

Without any regularization, this representation might end up representing a single light source with multiple point lights distributed over a sphere around the true source. To avoid such a scenario, we apply two regularization terms and an adaptive pruning scheme to motivate using the minimal number of point lights. We regularize the emission weight w_j of all SGs and penalize the inverse distance to the nearest surface d_{near} to move the lights further away from the reflections (Eq. (2)).

$$
\begin{aligned}
L_{pos} &= \sum_{i}^{N_{light}} 1/d_{i,near} \\
L_{val} &= \sum_{i}^{N_{light}} \sum_{j}^{N_{sg}} w_{i,j} \\
L &= L_{rec} + \lambda_{pos} L_{pos} + \lambda_{val} L_{val}
\end{aligned}
\tag{2}
$$

We use Adam optimizer [19] with initial learning rate $5e{-}2$, $\lambda_{pos}{=}1e{-}6$, $\lambda_{val}{=}1e{-}4$, $N_{light}{=}6{\times}8$, $N_{sg}{=}2{\times}6$. If the loss starts to stagnate, we reduce the learning rate by a factor of 0.5 and also prune the weakest light sources. We disable every light sources, which total intensity is smaller then 5% of the strongest light source. We stop the optimization if the performance stagnates longer. The whole optimization usually takes $5{-}10$ minutes on a single A6000 GPU depending on the scene complexity.

C. Experiment Details

Baselines. Both baselines [27, 48] have been trained for 320×240 resolution. As reported in the original papers, evaluating them on higher resolution leads to performance degradation; thus, we also evaluate them on this resolution.

D. Additional Results

D.1. Synthetic Results

We provide additional material estimation results on the InteriorVerse dataset [48] in Fig. 14.

Variance evaluation Single-view albedo estimation is an inherently ambiguous tasks, where specularity is one major source of ambiguity. We show the correlation between the metallic and albedo variance maps in Fig. 11. Glossy objects tend to have higher uncertainty, as also found quantitatively in the main text. Note that perfect correlation can not be expected, since specularity is not the only source of ambiguity.

D.2. Real Results

We provide additional material estimation results on the IIW dataset [4] in Fig. 15.

User study We conduct a user study to additionally evaluate the real-world predictions perceptually too.

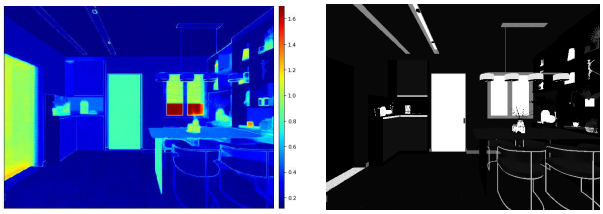
Image reconstruction We provide additional image rerenderings using our full pipeline in Fig. 12. We thank the authors of [48] for providing the code for running their method and for discussing the results.

S-AWARE Network [18] We compare against S-AWARE [18] in Fig. 13. Our method predicts roughness and metallic properties as well and improves upon the albedo estimation by avoiding baked-in lighting or shadows. We thank the authors for providing us with their results.

E. Additional Ablations

Effect of depth-conditioning. Our approach uses only a single image as input. However, geometry information can give beneficial cues for the appearance decomposition, since a physically-based renderer would require the normals for the shading and the depth for the global illumination estimation. To test this hypothesis, we train two other variants of our model. Both variants use additional depth and normal inputs. To provide a fair comparison between the variants, we evaluate all the methods without ground-truth depth and normal data. We use OmniData [11] to predict the depth and normal maps of the input view and use the predicted geometry as conditioning. One of our variants was trained with ground-truth geometry information, the other with predicted geometry.

We show qualitative results in Tab. 6. Here, we use the mean of 10 samples. Indeed, geometry information provides helpful cues for appearance decomposition and can improve the performance of our model. However, to stay consistent with all the other baselines, we kept the image-conditioned version as our main model.



Albedo Variance | Metallic

Figure 11. **Variance evaluation.** To investigate one source of ambiguity, we show a comparison between the variance of our albedo predictions and the true metallic map. Glossy objects tend to have higher variance due to the specular ambiguity. Further ambiguity arises e.g. from emissive objects, small, under- or over-exposed objects.

	InteriorVerse				IIW
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	WHDR ↓
Ours - ImageOnly	17.42 ± 3.08	0.80 ± 0.08	0.22 ± 0.08	25.42	22.02 ± 11.99
Ours - GTDepth	16.57 ± 3.79	0.80 ± 0.08	0.22 ± 0.08	24.36	17.05 ± 10.29
Ours - PredDepth	**18.31** ± 3.44	**0.82** ± 0.08	**0.19** ± 0.07	**22.60**	**16.66** ± 10.21

Table 6. **Effect of depth-conditioning.** Geometry information can give helpful cues for appearance decomposition and improve our models performance. We compare our image-conditioned method with additional conditioning on depth and normal maps. During training, one variant uses ground-truth geometry, the other uses predicted geometry using OmniData [11]. In test time, we use the predicted geometry as conditioning. Nevertheless, to stay consistent with the baselines, we kept the image-conditioning variant.

Figure 12. **Image reconstruction.** Additional image reconstruction results using our full pipeline.

Figure 13. **Comparison to S-AWARE [18].** Our method gives complex material maps and also improved albedo without baked-in lighting and shadows.

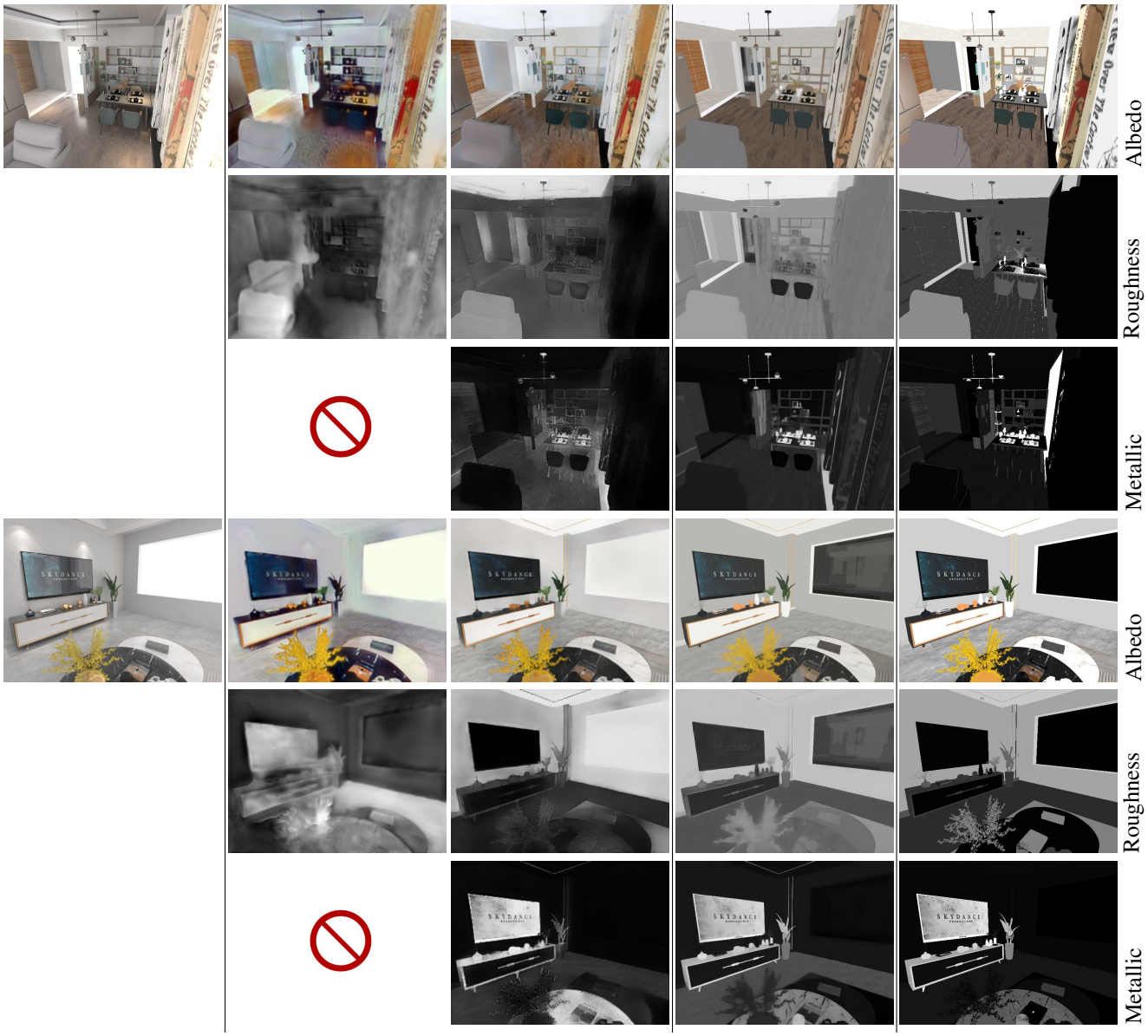


Figure 14. **Synthetic material estimation.** Continues on the next page.

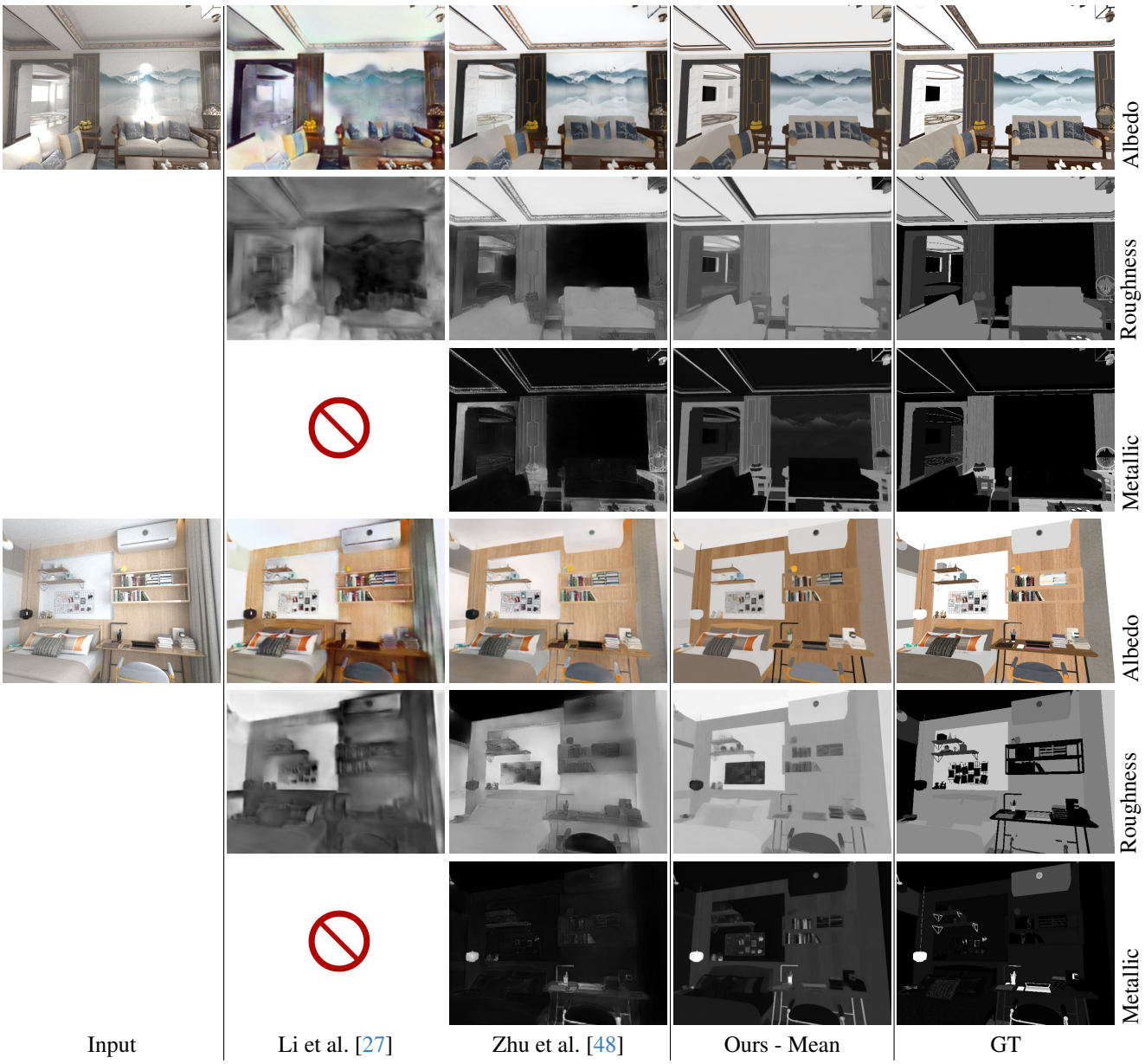


Figure 14. **Synthetic material estimation.** We compare our material estimation against the baselines [27, 48]. Both baselines produce good albedo colors overall, but they tend to bake in the lighting and specularities into the albedo map. In contrast, our method can produce clear materials with sharp edges and fine details.

Figure 15. **Real-world material estimation.** Continues on the next page.

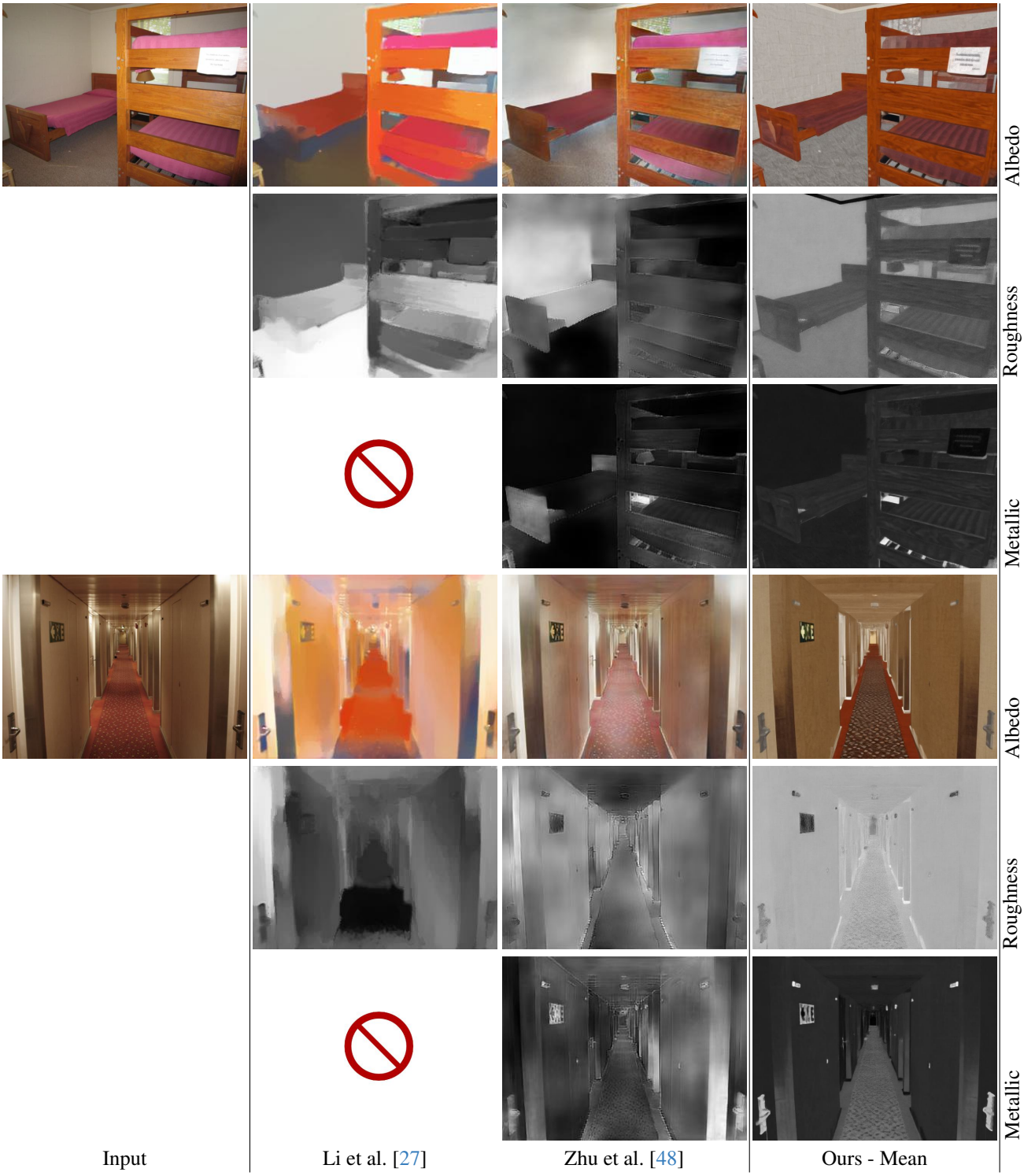


Figure 15. **Real-world material estimation.** We compare our material estimation against the baselines [27, 48]. Real-world lighting and shadows pose a bigger challenge for the baselines and they often bake them into the albedo map. Our method can produce sharp and detailed materials even in challenging real-world settings.