This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Intrinsic Image Diffusion for Indoor Single-view Material Estimation

Peter Kocsis
Technical University of Munich

Vincent Sitzmann
MIT EECS

Matthias Nießner
Technical University of Munich

peter-kocsis.github.io/IntrinsicImageDiffusion/

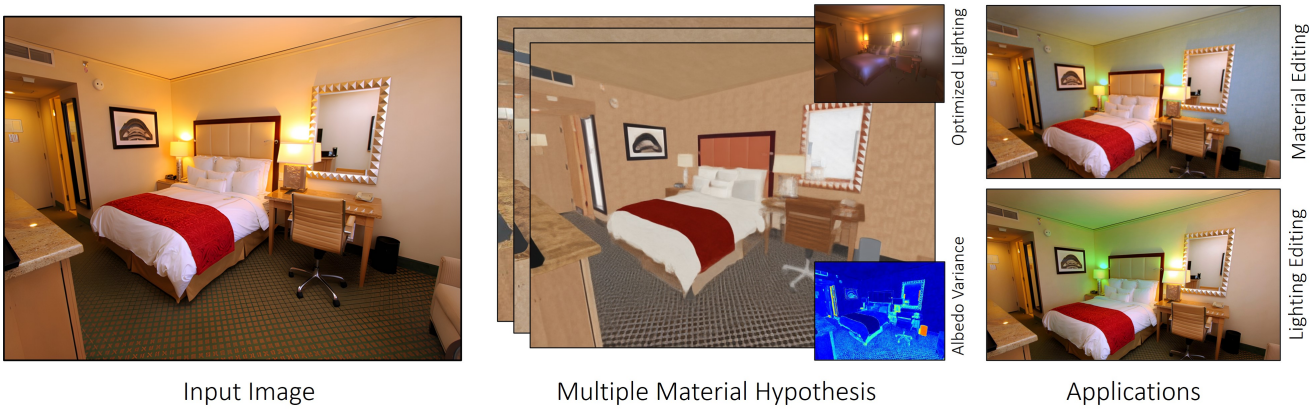

Figure 1. **Intrinsic Image Diffusion.** We present Intrinsic Image Diffusion for single-view material estimation of indoor scenes. Since appearance decomposition is a highly ambiguous task, we propose to use a conditional generative model to predict multiple solutions and utilize the strong prior of recent diffusion models [36]. Our approach gives detailed and consistent material estimations on complex indoor scenes, outperforming recent state-of-the-art methods and allows for high-quality controllable lighting optimization.

Abstract

We present Intrinsic Image Diffusion, a generative model for appearance decomposition of indoor scenes. Given a single input view, we sample multiple possible material explanations represented as albedo, roughness, and metallic maps. Appearance decomposition poses a considerable challenge in computer vision due to the inherent ambiguity between lighting and material properties and the lack of real datasets. To address this issue, we advocate for a probabilistic formulation, where instead of attempting to directly predict the true material properties, we employ a conditional generative model to sample from the solution space. Furthermore, we show that utilizing the strong learned prior of recent diffusion models trained on large-scale real-world images can be adapted to material estimation and highly improves the generalization to real images. Our method produces significantly sharper, more consistent, and more detailed materials, outperforming state-of-the-art methods by $1.5dB$ *on PSNR and by* 45% *better FID score on albedo prediction. We demonstrate the effectiveness of our approach through experiments on both synthetic and real-world datasets.*

1. Introduction

Intrinsic image decomposition is a long-standing challenge in computer vision. The goal is to predict geometric, material, and lighting properties from a single image, enabling numerous downstream applications, such as content editing by changing material or lighting conditions. However, the fundamental difficulty comes from the fact that we can only observe the visual appearance of objects as a result of a complex interplay of lighting and material, making the decomposition inherently ambiguous.

Recent data-driven algorithms [27, 48, 50] show significant improvement by utilizing large-scale synthetic datasets [27] rendered photo-realistically [48]. They achieve impressive results in predicting the overall material properties of objects in the scene; however, accurately capturing high-frequency details continues to pose a challenge.

We believe the limitations observed in existing methods can be attributed to their deterministic treatment of the appearance decomposition problem. Decomposing the appearance into lighting and material properties is a highly ambiguous task since lighting properties can be baked into materials and vice versa. By imposing the constraint of providing a single solution, deterministic models tend to predict the local or global average of the solution space, resulting in the aforementioned issues. We propose embracing the probabilistic nature of the problem, enabling a more

comprehensive exploration of the solution space.

In this work, we propose to formulate the task of single-view appearance decomposition as a probabilistic problem. We develop a generative model capable of sampling from the solution space conditioned on the input image. Specifically, we build upon the recent advancements of generative models and train a conditional diffusion model [15, 36] to predict the underlying material properties of a single image. With our approach, we aim to overcome the limitations of deterministic models and provide a more comprehensive and accurate representation of the underlying appearance decomposition problem.

Since no large-scale real-world intrinsic image decomposition dataset is available yet, previous methods usually train on synthetic data and may fine-tune on the real IIW dataset [4] to reduce the domain gap. Instead, we utilize the strong learned prior of diffusion models and fine-tune a pre-trained Stable Diffusion V2 [36] model to adapt its image prior to the task of material estimation using the photo-realistic synthetic InteriorVerse dataset of [48].

Our approach is not restricted to predicting a single solution; thus, it provides clear and sharp possible material explanations for a given observation without local or global averaging. Furthermore, by adapting the strong prior of pre-trained diffusion models, we aim to bring two key benefits. First, our objective is to reduce the domain gap caused by fine-tuning on synthetic imagery. Second, the prior of real images can already contain important cues that can help in material estimation, such as semantic and perceptual information [48]. In summary, our main contributions are:

1. We formulate the ambiguous appearance decomposition problem probabilistically and utilize diffusion models to sample solutions.
2. We leverage the real-world image prior of diffusion models for material estimation, achieving 77.6% FID and $4.04dB$ PSNR improvement in albedo prediction.
3. Using our material predictions, we optimize for lighting as multiple point lights and a global environment map.

2. Related work

Material estimation. The field of intrinsic image decomposition in computer graphics and vision has a long and extensive history, with research dating back to the 1970s. Earlier works aimed to decompose an image of a single object [14] into reflectance (albedo) and shading (irradiance) maps using heuristic approaches. Some of these methods include considering the image gradients [18, 22], employing chromaticity clustering [13], utilizing depth cues [8], or incorporating complex priors on the geometry, material, and lighting [2, 3]. [4, 21] took a significant step towards and introduced real-world datasets of indoor scenes for intrinsic image decomposition. These datasets include manually annotated relative comparisons of material properties, providing a valuable benchmark for evaluation. Additionally, [4] proposed a conditional-random-field-based algorithm demonstrating impressive improvement over prior methods.

Recent material estimation methods are usually part of a full inverse rendering framework [1, 9, 31, 32, 43, 49] and often use a learning-based algorithm for complex material predictions trained on increasingly photo-realistic datasets [27, 48] possibly combined with real data [24, 25]. There has been great improvement in the model architectures, including UNet-based [28, 43, 48], or cascaded networks [26, 27], and Transformers [50]. However, all of these methods use a feed-forward model to decompose the input image. We propose to formulate the task of appearance decomposition probabilistically and use a conditional generative model to sample multiple possible solutions.

Another line of work focuses on material acquisition, i.e. obtaining high-quality material properties of a single object. [35] introduced a similar idea for modeling the acquisition from a single scan probabilistically to address the disambiguity between specular and diffuse terms. Our method focuses on the single image multi-object material estimation and utilizes pre-trained image prior to better generalize.

Generative models. The field of generative architectures has recently gained significant attention with diffusion-based models [15, 20, 36, 39–41] contributing notable advancements to the field as highlighted in the surveys of [33, 45]. These models have demonstrated remarkable fidelity utilizing large-scale text-image datasets [38]. Moreover, these models possess a strong and robust prior that has proven valuable for various downstream applications across different domains [45], such as relighting [5] and even for extracting intrinsic properties [7, 10, 23] determinisically.

In our research, we leverage diffusion models to generate multiple samples from a solution space, harnessing their inherent probabilistic behavior. Building upon this, we propose to exploit the learned prior of recent diffusion models and transfer it to our specific task of material estimation. Similar to approaches such as [29, 47], we fine-tune StableDiffusion [36]. However, our unique contribution lies in targeting the material space instead of the image space. By adopting this approach, we aim to leverage the similarities between the material and image domains and utilize important cues for material estimation learned from real-world images, while simultaneously mitigating the domain gap.

3. Method

We aim to generate possible material properties $\hat{m} \in \mathbb{R}^{H\times W\times 5}$ consisting of albedo, roughness, and metallic from a single input image $x \in \mathbb{R}^{H\times W\times 3}$ ($\hat{m} \sim q(m|x)$). To this end, we train a latent diffusion model conditioned on the input image. We provide more details about the diffusion process in Sec. 3.2 and our material estimation in Sec. 3.1. We illustrate our training pipeline in Fig. 2.

Figure 2. **Training pipeline.** We train a conditional diffusion model to predict albedo and BRDF properties (roughness and metallic) given a single input image. We adapt the learned prior of Stable Diffusion [36] by fine-tuning it on the synthetic InteriorVerse [48] dataset. Models being trained are marked with yellow. (i) First, we separately encode the ground-truth (GT) albedo and BRDF properties with a fixed encoder to obtain the material feature maps. We also encode the conditioning image with a trainable encoder. (ii) We add noise to the material features and use our conditional diffusion model to predicted the noise. (iii) The training is supervised with L2 loss between the original and predicted noise. (iv) Using the predicted noise, the predicted material properties can be decoded separately.

Dataset We train our model on the photo-realistic InteriorVerse [48] dataset, which consists of more than 50000 rendered images with albedo, roughness, and metallic maps from over 4000 high-quality scenes. All the images and material maps are HDR in linear space.

3.1. Material representation

Estimating the material from a single image requires to decouple the material properties from the lighting. This task brings two main challenges. First, the appearance decomposition is highly ambiguous. Multiple joint material and lighting explanations can explain the same input view. The most obvious ambiguity is the scaling; the material can be scaled with an arbitrary global value, whereas the lighting can be downscaled accordingly. Furthermore, shading effects, such as shadows and specular highlights, can be baked into the material, resulting in valid explanations for different lighting conditions. Second, the material models used in computer graphics are hand-crafted approximations of real-world reflectance properties. Although they might be physically based, they don't exactly capture the real reflectance, leading to a representational gap.

In our work, we follow Zhu et al. [48] and represent the material with the physically-based GGX microfacet BRDF [42] consisting of albedo, roughness, and metallic properties. We create an image of the BRDF parameters, where channel R is the roughness, G is metallic, and B is always zero. To get the latent material features for the diffusion process, we separately encode the BRDF properties and the albedo and concatenate them, as shown in Fig. 2.

3.2. Training

Diffusion models have recently shown excellent capabilities in learning a prior over high-dimensional input data [15, 36]. These methods use an iterative denoising process to transform pure Gaussian noise into a data sample from the training distribution. Training them on large-scale datasets [38] gives them an extremely strong prior over real-world images.

Our goal is to utilize their learned prior and adjust it to the material estimation task. We fine-tune the pre-trained text-conditional Stable Diffusion V2 (SD) [36] on our task on synthetic data, similar to [29]. As [36], we follow the latent diffusion architecture with pre-trained frozen encoder $\mathcal{E}$ and decoder $\mathcal{D}$. To condition our diffusion model, we input the image information in two ways to the model. First, we encode it with a trained encoder $\mathcal{E}^*$ and concatenate it to the input noisy features. Our encoder $\mathcal{E}^*$ has the same architecture as the pre-trained encoder $\mathcal{E}$, but it is initialized randomly and gives 3-channel features; thus, the overall input channel size is 11. We found that naive downscaling to the latent dimension, as in the original depth-conditioned SD, causes loss in the high-frequency details. Using the frozen pre-trained encoder $\mathcal{E}$ performs worse, suggesting that the task of material estimation requires a different set of features for optimal predictions. Second, we use the CLIP [16, 34] image embedding as a cross-attention conditioner.

In each training step, we take a batch of images with the corresponding albedo and BRDF maps. We encode both material maps separately, then concatenate them in the la-

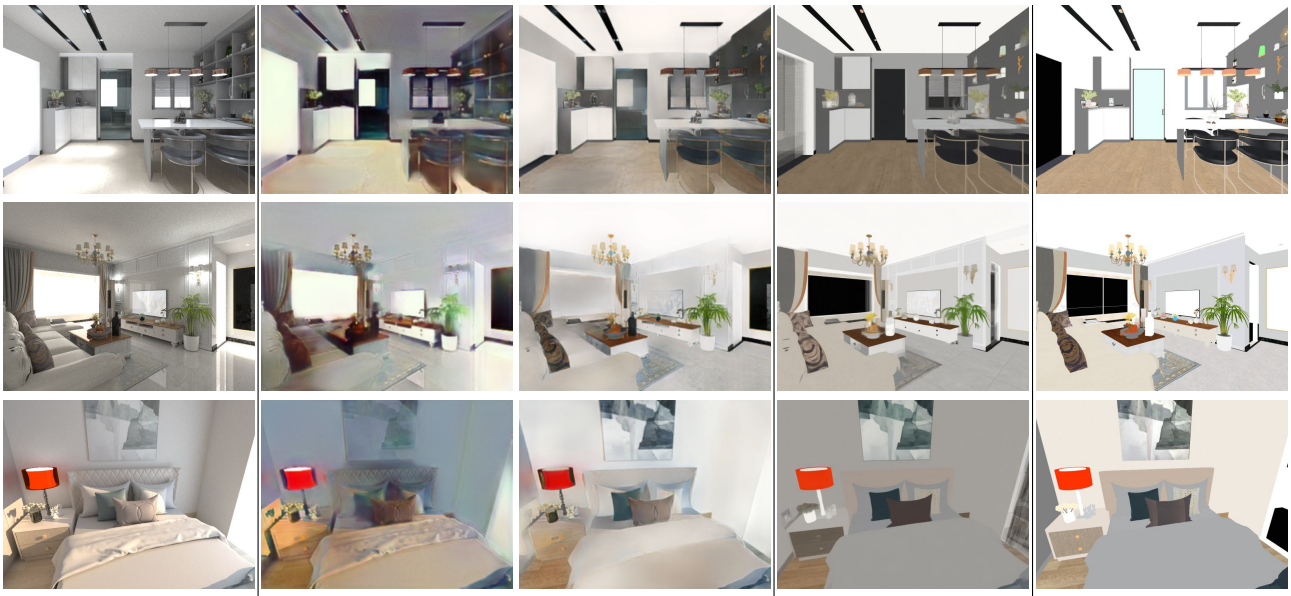

Figure 3. **Synthetic evaluation.** We qualitatively compare the predicted albedos of our to the baselines [27, 48] on the InteriorVerse dataset [48]. Both of the baselines produce smoothed results, often with baked-in lighting, specularities, or shadows. In contrast, our method gives sharp and detailed predictions with consistent textures. See supplementary for more results with roughness and metallic predictions.

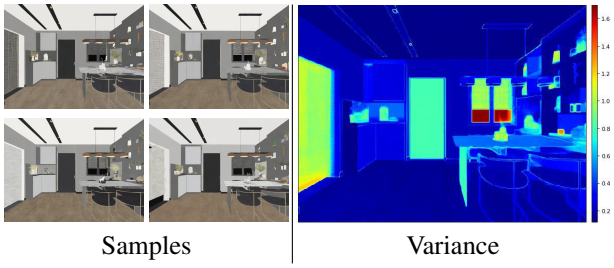

Figure 4. **Sample diversity.** We show multiple samples for a single scene and visualize the variance of the images across 100 samples. Specular and emissive objects have higher variance since their material properties are highly ambiguous.

tent space to get the material features. For each image, we sample a timestep $t \sim [1, 1000]$ and noise the material features accordingly [36]. We predict the added noise ϵ conditioned on the input image x. Our training objective is to minimize the distance between the original and the predicted noise, as shown on Fig. 2.

$$L = \mathbb{E}_{m,\epsilon\sim\mathcal{N}(0,I),t}[||\epsilon - \epsilon_\theta(\mathcal{E}(m) + \epsilon, t, x||_2^2] \quad (1)$$

In contrast to prior methods, our approach introduces a novel strategy for mitigating the domain gap. Prior work trains on synthetic data and subsequently fine-tunes on a smaller real-world dataset (IIW [4]) to adapt the synthetic prior to real-world scenarios. We take a different approach by fine-tuning a prior trained on large-scale real image data and adapting it to our specific task on synthetic data [29].

Figure 5. **Lighting Optimization.** We optimize for N_{light} point light sources with SG emission profile together with global incident lighting. Our representation is expressive enough to capture detailed emissions yet controllable for relighting purposes.

3.3. Inference

At test time, we use the regular diffusion process [36] to sample material features conditioned on a single image. We split the predicted material features into albedo and BRDF features and decode them separately to obtain a possible material explanation. We sample multiple possible solutions for one image and decode the albedo and BRDF predictions separately.

3.4. Lighting optimization

Given our high-fidelity, consistent material predictions, we optimize for the lighting of the scene. We use OmniData [11] to estimate the geometry. We optimize for global Spherical Gaussians (SG) [27] incident lighting and N_{light} point light sources with SG emission profile, resulting in a controllable yet expressive representation. We give more details in the supplemental. The optimization progress is shown in Fig. 5.

4. Experiments

In the following, we evaluate our method on both synthetic (Sec. 4.1) and real-world (Sec. 4.2) datasets. Furthermore, we ablate the effect of the key design choices of our method in Sec. 4.3. We show qualitative albedo prediction and quantitative material prediction (albedo, roughness, metallic) results. More qualitative results with additional ablation can be found in our supplementary material.

Experimental setting. We fine-tune the pre-trained Stable Diffusion V2 model [36] on the InteriorVerse dataset [48] with batch size of 40 for 250 epochs with a fixed learning rate of $1e$-5 using the AdamW optimizer [30]. To make our method robust against exposure changes, we normalize the input image to have a mean of 0.5 and clip them for the range $[0, 1]$. Furthermore, as expected by the pre-trained encoder $\mathcal{E}$, we transform all input maps to the range $[-1, 1]$. We use random crops of 256×256 resolution without any padding. The training takes ~ 6 days on 4 A6000 GPUs.

Although we train our model on 256×256 resolution crops, we evaluate our model on the full image resolution during inference, which is 640×480 for InteriorVerse [48] and $\sim 512 \times 384$ for IIW [4]. If not stated otherwise, we sample 10 possible materials for each input image. We use DDIM sampling [40] with 50 steps, which takes 17 seconds for 10 samples.

Baselines. We compare our method with [48] and [27], two recent state-of-the-art inverse rendering frameworks for indoor scenes. These methods use a UNet architecture [37] to deterministically predict geometric and material properties from a single image. Furthermore, we also report results for the optimization-based method of [4]; however, this method was tuned on the IIW dataset [4].

Metrics. Since we model the appearance decomposition probabilistically, we cannot directly compare our prediction to the GT. Therefore, we sample multiple samples for the same view and report the FID score of the predicted albedos alongside the mean of PSNR/SSIM/LPIPS metrics, the metrics of the mean sample, and the nearest sample. We calculate the mean sample in the image space.

Furthermore, albedo is a scale-invariant feature due to the inherent ambiguity between lighting and material intensity. Following [27], we report scale-invariant metrics for the albedo evaluations.

4.1. Synthetic results

We evaluate the effectiveness of our material estimation on the test set of the InteriorVerse dataset [48]. We provide various metrics to evaluate different aspects of our probabilistic model.

Fidelity. Our first experiment is about to show that our method predicts more similar albedos to the target albedo distribution. To assess the distribution similarity, we evaluate the FID score of the predictions in Tab. 1. [48] shows great improvement over [27] of 40.60% thanks to the training on the photo-realistic InteriorVerse dataset [48] and using an additional perceptual loss helping to utilize semantic information for sharper boundaries. Our method takes the next step and additionally improves over [48] by 44.99%.

	InteriorVerse			
	PSNR ↑	SSIM ↑	LPIPS ↓	FID
IIW [4]	9.73 ± 4.25	0.62 ± 0.14	0.47 ± 0.14	62.22
Li et al. [27]	12.31 ± 3.32	0.68 ± 0.11	0.52 ± 0.11	77.79
Zhu et al. [48]	15.92 ± 3.93	0.78 ± 0.09	0.34 ± 0.09	46.21
Ours - Mean metric	16.14 ± 3.27	0.73 ± 0.10	0.26 ± 0.11	—
Ours - Mean	17.42 ± 3.08	**0.80** ± 0.08	**0.22** ± 0.08	**25.42**
Ours - Best	**18.43** ± 2.97	0.77 ± 0.09	0.26 ± 0.11	—

Table 1. **Quantitative synthetic evaluation.** We quantitatively compare the predicted albedos between our method and the baselines [27, 48] on synthetic InteriorVerse data [48]. Additionally, we report the optimization-based method [4], which was directly tuned on the IIW dataset. We sample 10 solutions for each view and report the mean of the metrics, the metric of the mean, and the best prediction. (i) Our method has higher fidelity predictions closer to the target albedo distribution, as shown by the FID score. (ii) Our predictions are consistent with the input image, and we can produce predictions close to the GT since the best sample significantly outperforms all the others.

FID ↓	Albedo	Roughness	Metallic
Li et al. [27]	77.79	175.73	-
Zhu et al. [48]	46.21	69.92	48.71
Ours	**25.42**	**42.02**	**22.62**

Table 2. **Roughness and metallic evaluation.** We compare the quality of predicted roughness and metallic maps between our method and the baselines [27, 48] on the InteriorVerse dataset [48].

We show qualitative results in Fig. 3. [27, 48] gives reasonable albedo colors overall, but has difficulties with separating strong shadows, lighting, and specular effects from the material and often smoothes along the object boundaries. In turn, our approach produces smooth predictions over the same object but also maintains their high-frequency texture details maintaining sharp edges.

Consistency. In this section, we show that the predictions are consistent with the input image and it can be a valid solution for the input image. To this end, we evaluate in Tab. 1 how close the samples are to the GT albedo. We compare the PSNR, SSIM, and LPIPS metrics on the predicted albedos. Since appearance decomposition is a highly ambiguous task, a deterministic model [27, 48] yields just one sample from the solution space. Our approach can sample multiple solutions from the solution space. If we calcu-

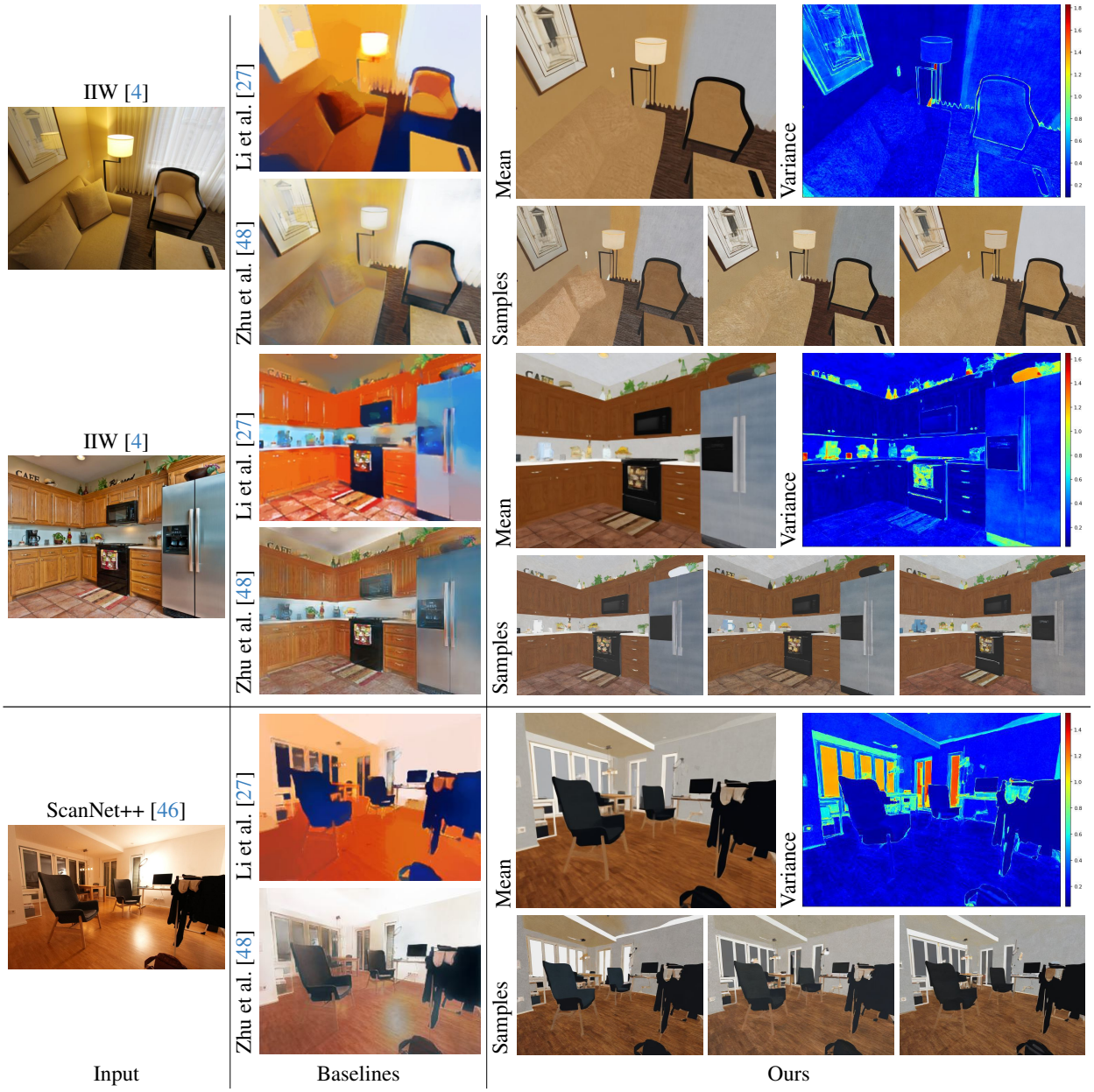

Figure 6. **Real-world evaluation.** We qualitatively compare the predicted albedos of our method to the baselines [27, 48] on the IIW [4] and ScanNet++ [46] datasets. Lights and shadows pose a challenge for the previous methods, but our approach gives consistent results on real-world inputs as well. We also visualize the variance of our predictions, showing how the specular, emissive, and small objects have higher uncertainty. See supplementary for more results with roughness and metallic predictions.

late the metric for all the samples independently and average them, we will get slightly better ($+0.22dB$ PSNR) but similar results as [48]. Taking the mean of the solution space drastically improves the performance ($+1.28dB$ PSNR), leading to the best SSIM. We also report the metric of the best samples too, showing that our solution space contains predictions much closer ($+1.01dB$ PSNR) to the ground truth. However, the ground truth is also just one sample in the solution space; thus, exactly finding it would require exploring this space.

Diversity. One remaining question about a generative model is whether it produces diverse samples without mode collapse. To evaluate this aspect, we sample 100 material predictions for a single view (top scene of Fig. 3). We show 4 predictions in Fig. 4 and visualize the relative variance of the samples. To this end, we normalize the predictions to zero mean and one standard deviation for each channel. Then, we plot the pixel-wise channel-average standard deviation. This variance map can be interpreted as an uncertainty map. We can see that large diffuse objects have stable

	IIW	
	WHDR ↓	PQ ↑
IIW [4]	21.00 ± 11.06	9.63
Li et al. [27]	21.99 ± 11.06	0.46
Zhu et al. [48]	22.90 ± 10.43	6.11
Ours - Mean metric	30.76 ± 14.05	—
Ours - Mean	22.02 ± 11.99	**83.80**
Ours - Best	**21.84** ± 11.63	—

Table 3. **Quantitative real evaluation.** We quantitatively compare the predicted albedos between our method and the baselines [27, 48] on real IIW data [4]. Additionally, we report the optimization-based method [4], which was directly tuned on the IIW dataset. We sample 10 solutions for each view and report the mean of the metrics, the metric of the mean, and the best prediction. We reach similar quantitative performance on the real-world IIW dataset; however, our predictions are much sharper and contain more high-frequency details (Figs. 6 and 7) as also assessed by Perceptual Quality in our user-study.

predictions with a small variance. However, the prediction of the specular, emissive, or small objects is more uncertain, showing complex ambiguity between lighting and material. To quantify this observation for specularity, we investigate the strength of the linear relationship between the albedo variance and metallic properties. We have evaluated the Pearson correlation value per image across all pixels for a subset of the test scenes, which ranges between 0.47 and 0.68, underlining our observation. Note that perfect correlation should not be expected as specularity is just one source of ambiguity. See qualitative comparison in supplemental.

Roughness and metallic evaluation. Next to the albedo predictions, we estimate per-pixel BRDF properties as well, such as roughness and metallic. We evaluate the fidelity of these predictions in Tab. 2. [27] does not estimate the metallic properties. Similar to the advancements observed in albedo predictions, [48] demonstrates significant improvement compared to [27], achieving a remarkable relative enhancement of 60.21% FID in roughness prediction. Building upon this achievement, our approach surpasses the performance of [48] by further improving 39.90% and 53.56% FID in roughness and metallic prediction, respectively. We provide qualitative results in our supplementary.

4.2. Real-world results

Qualitative evaluation. We evaluate our method on the challenging IIW [4] and ScanNet++ [46] datasets. Li et al. [27] proposes an additional bilateral solver for refinement and found that it helps for real-world images. This solver smoothes the material, making them more consistent overall but for the cost of losing high-frequency details. Following Li et al. [27], we use the bilateral solver for evaluating their method on real scenes.

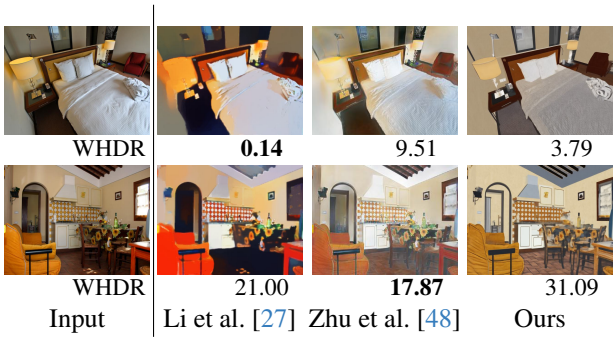

Figure 7. **WHDR metric imperfection.** The WHDR metric tends to favor smoother outcomes, thus not providing an optimal evaluation of high-frequency details. Despite our method producing significantly clearer and sharper predictions, the baseline achieves a better score even though their predictions exhibit burned-in lighting and over-smoothed texture.

We show qualitative results in Fig. 6. [27] often produces over-smoothed predictions but with overall correct colors. [48] gives more detailed albedos but still struggles with baked-in shadows, lighting, and specularity, especially close to light sources. Our method predicts detailed and well-textured materials over the objects and also for real-world samples. Similarly to the synthetic results, small, specular, and emissive objects have higher variance. For roughness and metallic results, see supplemental.

Quantitative evaluation. We report the WHDR metric [4] in Tab. 3, which compares the predicted albedo to human judgments of relative intensities. Our method achieves similar performance quantitatively as the baselines. However, the WHDR metric assesses predictions sparsely and tends to favor smoother outcomes, thus not providing an optimal evaluation of high-frequency details, as shown in Fig. 7 and also found in [6, 12, 44].

As an additional quantitative evaluation, we have conducted a user-study with 27 participants on assessing the albedo quality. We show the albedo prediction of three baselines [4, 27, 48] and our method together with the input image for a total of 40 scenes of the IIW dataset [4]. The user needs to choose the best prediction. To get the users more acquainted with assessing the quality of albedo predictions, we first show them three synthetic images with their ground-truth albedo. We report the percentage of being chosen as *Perceptual Quality* in Tab. 3.

Image Reconstruction. We evaluate the image reconstruction quality of our method. First, we use the lighting and geometry prediction module of [48] with our material predictions, which makes the renderings more consistent; however, fails to capture small light sources (Fig. 8). Our lighting optimization reconstructs all light sources faithfully outperforming the baseline with a high margin (Tab. 4)

Figure 8. **Image reconstruction.** We compare the rerendering of [48] on real images [4] first using their, then our material predictions against our full inverse rendering pipeline. Being trained on synthetic data, [48] suffers from domain gap on real-world samples, while our material predictions help to maintain sharp details, enabling to fit controllable detailed lighting.

	PSNR ↑	SSIM ↑	LPIPS ↓
Zhu et al. [48]	13.54 ± 2.22	0.51 ± 0.12	0.43 ± 0.08
[48] w/ Ours	14.07 ± 1.91	0.55 ± 0.11	0.38 ± 0.08
Ours Full	**21.96** ± 3.22	**0.70** ± 0.10	**0.22** ± 0.08

Table 4. **Image reconstruction.** We compare the image rerendering of [48] using their and our material predictions against our full pipeline with lighting estimation (Fig. 5). Our material predictions generalize better to real-world images thanks to the diffusion prior and improve the rendering of [48], even though their lighting estimation network is trained with predicted materials of [48] and forced to learn to compensate for their discrepancies. Our consistent materials enable us to fit detailed and also controllable lighting, improving the rerendering quality by a high margin.

4.3. Ablations

Effect of pre-training. Our main design choice is the fine-tuning of a pre-trained diffusion model. We evaluate the albedo predictions on the InteriorVerse dataset [48] and on the IIW dataset [4]. Both qualitatively Fig. 9 and quantitatively (Tab. 5), using the pre-trained model, gives significant improvement. Notably, pre-training not only aids in reducing the domain gap but also proves beneficial in synthetic scenarios, showing that the prior knowledge obtained from real imagery can be effectively adapted and leveraged for material estimation. Without pre-training, the model struggles with under and over-saturated regions, but pre-training gives a strong prior over various objects, enabling high-quality material estimation in challenging regions.

4.4. Limitations and Future Work

We have shown significant improvement for single-view material estimation by formulating the task probabilistically and utilizing the strong learned prior of a pre-trained diffusion model. Currently, we train on paired synthetic samples and use the real-world image prior of a pre-trained diffusion model to reduce the domain gap. Extending this method to weak supervision, even with real-world samples, is a great avenue for future research. Furthermore, our current work focuses on material estimation and optimizes for the lighting independently. A fused inverse rendering framework that effectively leverages the probabilistic nature of our method for lighting estimation could be beneficial. Finally, generative models present numerous innovative possibilities that offer greater control over renderings, such as text-guided material editing.

Figure 9. **Model ablation.** Leveraging Stable Diffusion's [36] real image prior significantly improves our material diffusion.

	InteriorVerse				IIW
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	WHDR ↓
Pix2Pix [17]	13.69 ± 2.94	0.67 ± 0.11	0.50 ± 0.11	84.28	36.42 ± 12.28
Ours w/o PT - Mean	13.38 ± 2.44	0.69 ± 0.09	0.42 ± 0.10	113.31	35.60 ± 12.94
Ours w/o PT - Best	13.85 ± 2.46	0.61 ± 0.12	0.61 ± 0.14		36.34 ± 11.78
Ours - Mean	17.42 ± 3.08	**0.80** ± 0.08	**0.22** ± 0.08	**25.42**	22.02 ± 11.99
Ours - Best	**18.43** ± 2.97	0.77 ± 0.09	0.26 ± 0.11		**21.84** ± 11.63

Table 5. **Model ablation.** Using a pre-trained diffusion model brings substantial improvement. Using the GAN-based pix2pix [17] performs slightly weaker as diffusion trained from scratch. Stable Diffusion's [36] prior of real imagery not only reduces the domain gap but can also be adapted for material estimation, resulting in improved synthetic results as well.

5. Conclusion

We have introduced Intrinsic Image Diffusion, a generative model for material estimation. Rather than predicting a single solution, we propose probabilistically formulating the highly ambiguous appearance decomposition. We have shown that leveraging the strong image prior of recent diffusion models improves the material prediction quality and reduces the domain gap. Our material predictions are significantly cleaner and more consistent than the ones of prior works, enabling direct optimization-based emissive lighting fitting. We believe our work can pave the way for novel approaches to tackling the challenging task of the ill-posed appearance decomposition by applying conditional generative models and utilizing image prior.

Acknowledgements. This work was supported by the ERC Starting Grant Scan2CAD (804724), the German Research Foundation (DFG) Grant "Making Machine Learning on Static and Dynamic 3D Data Practical", and the German Research Foundation (DFG) Research Unit "Learning and Simulation in Visual Computing". Vincent Sitzmann was supported by the National Science Foundation under Grant No. 2211259, and by the Singapore DSTA under DST00OECI20300823 (New Representations for Vision).

References

[1] Dejan Azinovic, Tzu-Mao Li, Anton Kaplanyan, and Matthias Nießner. Inverse path tracing for joint material and lighting estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2447–2456. Computer Vision Foundation / IEEE, 2019. 2

[2] Jonathan T. Barron and Jitendra Malik. Intrinsic scene properties from a single RGB-D image. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 17–24. IEEE Computer Society, 2013. 2

[3] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(8):1670–1687, 2015. 2

[4] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Trans. Graph.*, 33(4):159:1–159:12, 2014. 2, 4, 5, 6, 7, 8

[5] Anand Bhattad and D.A. Forsyth. Stylitgan: Prompting stylegan to generate new illumination conditions. In *arXiv*, 2023. 2

[6] Anand Bhattad and David A. Forsyth. Cut-and-paste object insertion by enabling deep image prior for reshading. In *International Conference on 3D Vision, 3DV 2022, Prague, Czech Republic, September 12-16, 2022*, pages 332–341. IEEE, 2022. 7

[7] Anand Bhattad, Daniel McKee, Derek Hoiem, and David A. Forsyth. Stylegan knows normal, depth, albedo, and more. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 2

[8] Qifeng Chen and Vladlen Koltun. A simple model for intrinsic image decomposition with depth cues. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 241–248. IEEE Computer Society, 2013. 2

[9] 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. *CoRR*, abs/2303.12368, 2023. 2

[10] Xiaodan Du, Nicholas I. Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let's find out! *CoRR*, abs/2311.17137, 2023. 2

[11] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10766–10776. IEEE, 2021. 4, 1, 2

[12] David A. Forsyth and Jason J. Rock. Intrinsic image decomposition using paradigms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7624–7637, 2022. 7

[13] Elena Garces, Adolfo Munoz, Jorge Lopez-Moreno, and Diego Gutierrez. Intrinsic images by clustering. In *Computer graphics forum*, pages 1415–1424. Wiley Online Library, 2012. 2

[14] Roger B. Grosse, Micah K. Johnson, Edward H. Adelson, and William T. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 2335–2342. IEEE Computer Society, 2009. 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 3

[16] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 3

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 8

[18] Yeying Jin, Ruoteng Li, Wenhan Yang, and Robby T Tan. Estimating reflectance layer from a single image: Integrating reflectance guidance and shadow/specular aware learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1069–1077, 2023. 2, 3

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2

[20] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *CoRR*, abs/2107.00630, 2021. 2

[21] Balazs Kovacs, Sean Bell, Noah Snavely, and Kavita Bala. Shading annotations in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 850–859. IEEE Computer Society, 2017. 2

[22] Edwin Land and John McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61:1–11, 1971. 2

[23] Hsin-Ying Lee, Hung-Yu Tseng, Hsin-Ying Lee, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable pixel-level semantic prediction. *CoRR*, abs/2311.18832, 2023. 2

[24] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, pages 381–399. Springer, 2018. 2

[25] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9039–9048. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[26] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Trans. Graph.*, 37(6):269, 2018. 2

[27] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 2472–2481. Computer Vision Foundation / IEEE, 2020. 1, 2, 4, 5, 6, 7

[28] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Milos Hasan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-based editing of indoor scene lighting from a single image. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, pages 555–572. Springer, 2022. 2

[29] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *CoRR*, abs/2303.11328, 2023. 2, 3, 4

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 5

[31] Merlin Nimier-David, Zhao Dong, Wenzel Jakob, and Anton Kaplanyan. Material and lighting reconstruction for complex indoor scenes with texture-space differentiable rendering. In *32nd Eurographics Symposium on Rendering, EGSR 2021 - Digital Library Only Track, Saarbrücken, Germany, June 29 - July 2, 2021*, pages 73–84. Eurographics Association, 2021. 2

[32] Julien Philip, Sébastien Morgenthaler, Michaël Gharbi, and George Drettakis. Free-viewpoint indoor neural relighting from multi-view stereo. *ACM Trans. Graph.*, 40(5):194:1–194:18, 2021. 2

[33] Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T. Barron, Amit H. Bermano, Eric Ryan Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, C. Karen Liu, Lingjie Liu, Ben Mildenhall, Matthias Nießner, Björn Ommer, Christian Theobalt, Peter Wonka, and Gordon Wetzstein. State of the art on diffusion models for visual computing. *CoRR*, abs/2310.07204, 2023. 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[35] Carlos Rodríguez-Pardo, Henar Dominguez-Elvira, David Pascual-Hernández, and Elena Garces. Umat: Uncertainty-aware single image high resolution material capture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 5764–5774. IEEE, 2023. 2

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. 1, 2, 3, 4, 5, 8

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, pages 234–241. Springer, 2015. 5

[38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 2, 3

[39] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2256–2265. JMLR.org, 2015. 2

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 5

[41] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *CoRR*, abs/2011.13456, 2020. 2

[42] Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. Microfacet models for refraction through rough surfaces. In *Proceedings of the Eurographics Symposium on Rendering Techniques, Grenoble, France, 2007*, pages 195–206. Eurographics Association, 2007. 3

[43] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning indoor inverse rendering with 3d spatially-varying lighting. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12518–12527. IEEE, 2021. 2

[44] Jiaye Wu, Sanjoy Chowdhury, Hariharmano Shanmugaraja, David Jacobs, and Soumyadip Sengupta. Measured albedo in the wild: Filling the gap in intrinsics evaluation. *COPR*, 2023. 7

[45] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Ming-Hsuan Yang, and Bin Cui. Diffusion models: A comprehensive survey of methods and applications. *CoRR*, abs/2209.00796, 2022. 2

[46] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 6, 7

[47] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *CoRR*, abs/2302.05543, 2023. 2

[48] 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, SA 2022,*

Daegu, Republic of Korea, December 6-9, 2022, pages 6:1–6:8. ACM, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[49] Jingsen Zhu, Yuchi Huo, Qi Ye, Fujun Luan, Jifan Li, Dianbing Xi, Lisha Wang, Rui Tang, Wei Hua, Hujun Bao, and Rui Wang. I^2-sdf: Intrinsic indoor scene reconstruction and editing via raytracing in neural sdfs. *CoRR*, abs/2303.07634, 2023. 2

[50] Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. Irisformer: Dense vision transformers for single-image inverse rendering in indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 2812–2821. IEEE, 2022. 1, 2