IntrinsicDiffusion: Joint Intrinsic Layers from Latent Diffusion Models

Jundan Luo
jundanluo22@gmail.com
University of Bath
United Kingdom, Bath

Duygu Ceylan
ceylan@adobe.com
Adobe Research
United Kingdom, London

Jae Shin Yoon
jaeyoon@adobe.com
Adobe Research
United States, San Jose

Nanxuan Zhao
nanxuanzhao@gmail.com
Adobe Research
United States, San Jose

Julien Philip
julienov.philip@gmail.com
Adobe Research
United Kingdom, London

Anna Frühstück
a.fruehstueck@gmail.com
Adobe Research
United Kingdom, London

Wenbin Li
wenbin.li@bath.edu
University of Bath
United Kingdom, Bath

Christian Richardt
christian@richardt.name
Codec Avatars Lab, Meta Reality Labs
United States, Pittsburgh

Tuanfeng Y. Wang
yangtwan@adobe.com
Adobe Research
United Kingdom, London

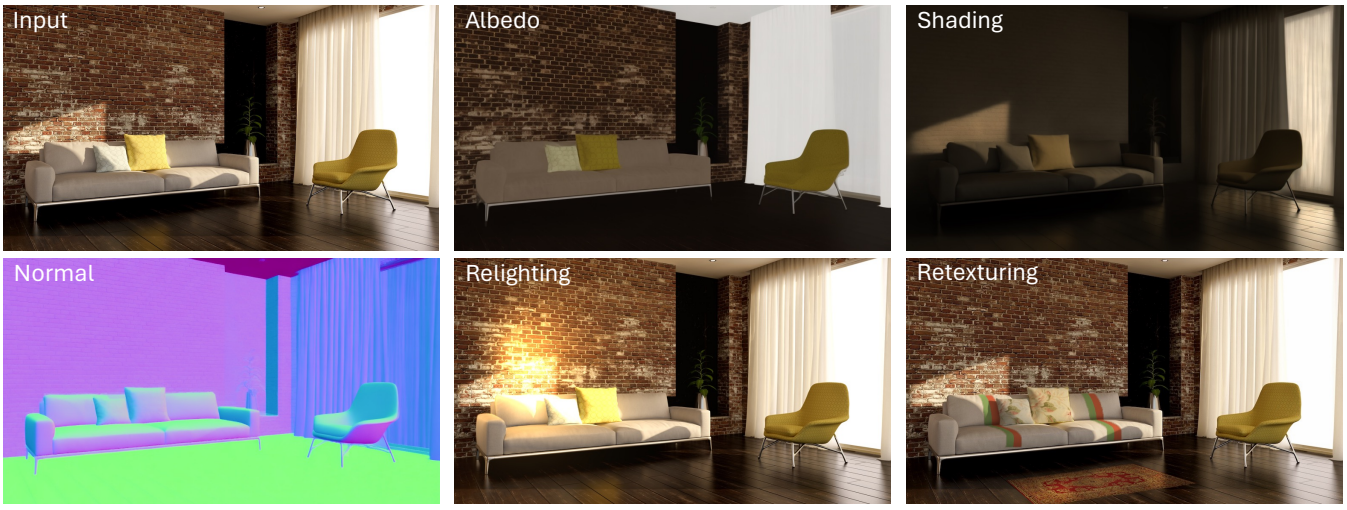

Figure 1: Results from our intrinsic image decomposition method that can jointly predict albedo, shading, and surface normal layers from a single image. These decomposed layers enable various editing tasks, such as image relighting and retexturing. Source: 3DarcaStudio, stock.adobe.com.

ABSTRACT

Reasoning about the intrinsic properties of an image, such as albedo, illumination, and surface geometry, is a long-standing problem with many applications in image editing and compositing. Existing solutions to this ill-posed problem either heavily rely on manually designed priors or learn priors from limited datasets that lack diversity. Hence, they fall short in generalizing to in-the-wild test scenarios. In this paper, we show that a large-scale text-to-image generation model trained on a massive amount of visual data can implicitly learn intrinsic image priors. In particular, we introduce a novel conditioning mechanism built on top of a pre-trained foundational image generation model to jointly predict multiple intrinsic modalities from an input image. We demonstrate that predicting different modalities in a collaborative manner improves the overall quality. This design also enables mixing datasets with annotations of only a subset of the modalities during training, contributing to the generalizability of our approach. Our method achieves state-of-the-art performance in intrinsic image decomposition, both qualitatively and quantitatively. We also demonstrate downstream image editing applications, such as relighting and retexturing.

SIGGRAPH Conference Papers '24, July 27–August 01, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07
https://doi.org/10.1145/3641519.3657472

CCS CONCEPTS

• **Computing methodologies** → *Scene understanding*; **Image processing**; **Image representations**.

KEYWORDS

intrinsic image decomposition, diffusion model, multi-task learning, surface normal estimation

ACM Reference Format:
Jundan Luo, Duygu Ceylan, Jae Shin Yoon, Nanxuan Zhao, Julien Philip, Anna Frühstück, Wenbin Li, Christian Richardt, and Tuanfeng Y. Wang. 2024. IntrinsicDiffusion: Joint Intrinsic Layers from Latent Diffusion Models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27–August 01, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3641519.3657472

1 INTRODUCTION

Understanding the intrinsic layers of an image reveals a scene's inherent properties. A typical intrinsic decomposition of an image **I** can be represented as (i) the reflectivity of the surfaces depicted in the image, denoted as the albedo layer **A**, (ii) the interaction between the light and the surfaces, denoted as the shading layer **S**, and (iii) the direct underlying surface geometry, denoted by the normal layer **N**. Understanding such properties from a single image is useful for various editing tasks, such as relighting [Basri and Jacobs 2003], retexturing [Jafarian et al. 2023; Ye et al. 2023; Zheng et al. 2022], and realistic image composition [Careaga and Aksoy 2023]. As such, intrinsic image decomposition has been one of the most fundamental tasks in computer graphics [Barrow and Tenenbaum 1978], and the pursuit of accuracy has lasted for decades. In this paper, we introduce a method that can jointly predict the intrinsic layers of albedo, shading, and surface normal from a single image, as shown in Figure 1.

Predicting the intrinsic properties from a single image is challenging for the following reasons. First, the same image can be represented by different combinations of intrinsic layers, making this problem a highly ill-posed task. Existing solutions either use manually crafted regularization priors or learn such priors from data to tackle this problem in a deterministic fashion. Second, existing scene-level datasets used for this task are either synthetic or only sparsely annotated since annotating intrinsic layers for real images is not straightforward. Developing a solution that can generalize to in-the-wild test cases is therefore not easy.

In the meantime, we are witnessing a revolution in large-scale text-to-image diffusion models in terms of the quality and the diversity of the images they can generate [Rombach et al. 2022]. Being trained on massive amounts of data, such foundational models learn to understand the implicit properties of scenes, such as different lighting conditions and appearance changes as shown in Figure 2. Such an understanding is critical to learning a general and effective prior. Moreover, being generative, such a model is suitable to tackle the ambiguous nature of the intrinsic image decomposition task. Hence, the main question we ask is "How to effectively repurpose a pre-trained foundational image generation model for our task?"

The main difficulty in direct finetuning or continued training of the foundational model for our task stems from the fact that the amount of training data available for intrinsic image decomposition with ground-truth labels is orders of magnitude smaller than the data used to train the text-to-image model. The recent ControlNet architecture [Zhang et al. 2023] tackles a similar problem for effective conditioning of the image generation model on various control signals using datasets of moderate size. Inspired by this success, we propose a novel approach that utilizes ControlNet to repurpose the text-to-image generation model for the intrinsic image decomposition task. Given an input image as the condition to a control model, our goal is to generate the corresponding intrinsic layers, i.e., albedo, shading, and surface normals. Instead of training a separate control model for each layer, the core of our approach is a joint control branch that adapts to each modality via different prompts. We highlight that our multimodal learning framework allows the network to be trained on assorted datasets with different types of available annotations. For example, Vasiljevic et al. [2019] provide images with accurate surface normal annotations only. The ability to mix such datasets addresses the dataset problem to a large extent. We also show that this framework enables learning a more effective latent space where the prediction of one modality (e.g., surface normals) helps to improve the prediction of other modalities (e.g., albedo and shading, as shown in Figure 6 and Table 3).

Figure 2: We generate images of a room using a pre-trained latent diffusion model with different lighting condition prompts. As shown, the model has implicit knowledge of the intrinsic image properties of the scene.

To improve the spatial alignment between the generated intrinsic layers as well as to prevent spatial distortion artifacts, we further upgrade the conditional image encoder used in ControlNet to a sequence of residual blocks [He et al. 2016] continued with SwinV2 transformer layers [Liu et al. 2022]. While the weights of the residual blocks are shared between different modalities, we use separate transformers to map the intermediate features to each modality.

We evaluate our method on both synthetic and real datasets. We perform comparisons to previous approaches and show state-of-the-art performance on various challenging cases where our method shows strength in generalizability and visual quality, together with competitive quantitative performance on benchmarks. We also carefully ablate the different design choices we adapt. Finally, we show various downstream applications such as retexturing and relighting that benefit from our method.

In summary, our main contributions include:

- casting the intrinsic image decomposition as a conditional generation problem that leverages a pre-trained foundational text-to-image model;
- a novel ControlNet architecture that jointly predicts multiple intrinsic modalities (e.g., shading, albedo, and surface normal) and achieves state-of-the-art performance;

- the ability to combine different data sources with different types of annotations through a joint learning framework, improving the overall performance of our method.

2 RELATED WORK

Intrinsic image decomposition is a classic visual computing task that has been approached in many different ways. Traditional approaches focus on using optimization while learning-based approaches took the lead more recently. Comprehensive surveys by Bonneel et al. [2017] and Garces et al. [2022] provide a good overview.

2.1 Optimization-based Intrinsic Image Decomposition

To tackle the ill-posed intrinsic image decomposition problem, optimization-based approaches rely on various priors and assumptions. This includes smooth shading [Chen and Koltun 2013; Garces et al. 2012], grayscale shading [Garces et al. 2012; Grosse et al. 2009; Zhao et al. 2012], and albedo sparsity [Bell et al. 2014; Bi et al. 2015; Garces et al. 2012; Gehler et al. 2011; Meka et al. 2021; Shen et al. 2011], among other things. Several approaches also explore additional information for disambiguation, such as 3D geometry [Chen and Koltun 2013; Hachama et al. 2015; Meka et al. 2017; Wu et al. 2014; Yu et al. 2013; Zollhöfer et al. 2015] or user interaction [Bousseau et al. 2009; Meka et al. 2017; Shen et al. 2011]. Recently, several works have also tackled the intrinsic decomposition problem in the context of radiance fields [Choi et al. 2023; Sarkar et al. 2023; Ye et al. 2023; Zhu et al. 2023], where multi-view images are available. Such optimization methods are also used to aid editing tasks like illumination editing [Huang et al. 2023; Shah et al. 2023]. While manually crafted priors have been effective, more recently, the trend has been to learn these priors from data, as we discuss next.

2.2 Learning-based Intrinsic Image Decomposition

Neural networks can learn priors implicitly given sufficient training data. An early key catalyst was the IIW dataset [Bell et al. 2014], which provides sparse ordinal annotations on real-world "intrinsic images in the wild" for darker/similar/brighter albedo values. This dataset enabled the first wave of approaches that predict relative reflective relationships [Narihira et al. 2015; Zhou et al. 2015; Zoran et al. 2015]. The CGIntrinsics (CGI) dataset [Li and Snavely 2018a] was the first major dataset providing per-pixel ground-truth albedo thanks to its synthetic nature. Since then, many convolutional neural networks were proposed [Fan et al. 2018; Jin et al. 2023; Li and Snavely 2018a; Liu et al. 2020; Wang et al. 2023]. More recent datasets like OpenRooms [Li et al. 2021b], Hypersim [Roberts et al. 2021] and InteriorVerse [Zhu et al. 2022] further increase the visual fidelity of synthetic data and provide high dynamic range images. In pursuit of additional constraints, some methods have also explored multi-task training via models that jointly learn depth and/or normal prediction with intrinsic image decomposition [Kim et al. 2016; Luo et al. 2020; Zhou et al. 2019]. Other approaches learn in an unsupervised fashion from multi-illumination datasets [Careaga and Aksoy 2023; Lettry et al. 2018; Li and Snavely 2018b]. Finally, several works [Das et al. 2023; Forsyth and Rock 2022; Luo et al. 2023] have explored the local characteristics of the intrinsic image decomposition problem, i.e., albedo predictions in overlapping local patches should be consistent to provide additional supervision signal. Empowered by a pre-trained foundational model, our method excels at capturing intricate details. We eliminate the priors and assumptions employed by previous methods for the sake of simplification, enabling our model to be fully exposed to complex real-world data.

2.3 Generative Models and Conditional Image Generation

Generative models have established themselves as the widespread solution for high-quality image synthesis. Following the success of StyleGAN [Karras et al. 2021], more recently, diffusion models [Ho et al. 2020; Rombach et al. 2022] have dominated image generation and editing tasks. We refer the reader to detailed surveys for both GANs [Bermano et al. 2022] and diffusion models [Po and Wetzstein 2023; Yang et al. 2023]. The generative nature of such models has been shown to be useful to tackle various image analysis tasks such as semantic segmentation [Li et al. 2021a], depth estimation [Saxena et al. 2023a,b], and intrinsic image decomposition [Shah et al. 2023]. More recently, there have been concurrent efforts to leverage pre-trained image generation models for the intrinsic image decomposition task. Bhattad et al. [2023] perform such an analysis in the StyleGAN space. Since recent text-to-image diffusion models leverage datasets that are orders of magnitude larger than used by GANs, our method instead utilizes a latent diffusion model for a similar task. The concurrent work of Kocsis et al. [2024] finetunes stable diffusion by treating the intrinsic layers as a multi-channel image. Hence, their method requires access to datasets that provide annotations for all the layers during training. In a similar fashion, Du et al. [2023] train a LoRA adaptor [Hu et al. 2022] to predict each modality. In contrast, our work presents a multimodal training strategy with a joint image encoder that enables the mixing of datasets with different types of available annotations.

3 METHODOLOGY

Given an input image $\mathbf{I} \in \mathbb{R}^{W \times H \times 3}$, our goal is to generate the corresponding intrinsic layers. Specifically, we generate an albedo layer $\mathbf{A} \in \mathbb{R}^{W \times H \times 3}$, a colored shading layer $\mathbf{S} \in \mathbb{R}^{W \times H \times 3}$, and a surface normal map $\mathbf{N} \in \mathbb{R}^{W \times H \times 3}$. To this end, we first extract embedding features from the input image $\mathbf{I}$ (Section 3.2). These features are then mapped to domain-specific embedding vectors that are used as conditioning input for a conditional latent diffusion model via a ControlNet [Zhang et al. 2023] (Section 3.3). To adapt the network to different intrinsic modalities (e.g., albedo, shading, and surface normal), we jointly learn a set of shared residual blocks and domain-specific transformers for feature embedding, and use domain-specific text prompts to condition the ControlNet simultaneously. We illustrate our system pipeline in Figure 3 and discuss the details in the following sections.

3.1 Preliminaries

Diffusion models [Ho et al. 2020; Sohl-Dickstein et al. 2015] consist of a forward and a backward process. The forward process adds

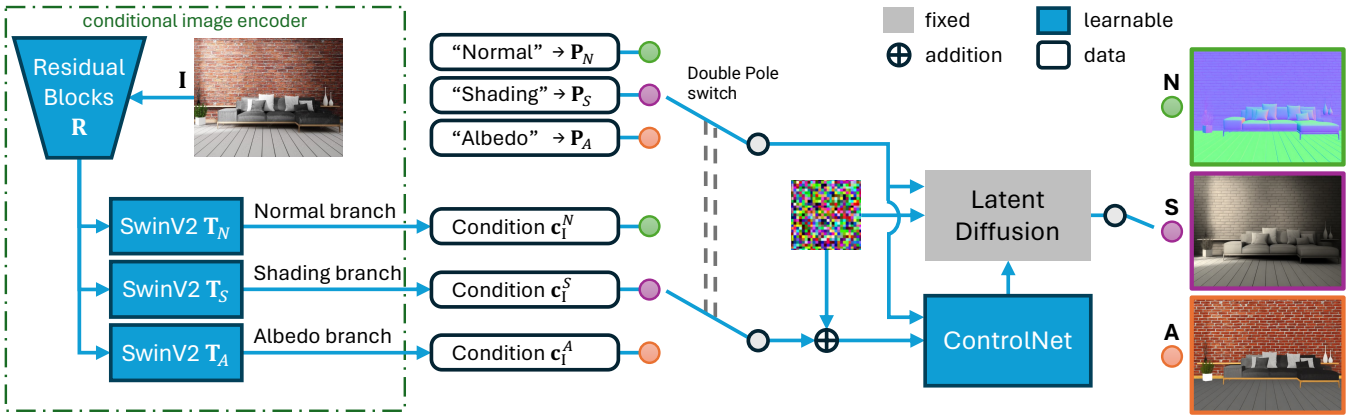

Figure 3: Overview of our pipeline. Given an input image, we extract image features using a conditional image encoder with residual blocks. The image features are projected to domain-specific conditional vectors using SwinV2 [Liu et al. 2022] blocks for each intrinsic modality, i.e., albedo, shading, and surface normal (Section 3.2). The generative diffusion model with a ControlNet [Zhang et al. 2023] is trained to generate different intrinsic modalities, and acts as a function of the dedicated prompt and conditional vectors (Section 3.3). Image source: Interior Design, stock.adobe.com.

Gaussian noise to the data to gradually remove information. In latent diffusion models [Rombach et al. 2022], the data is first mapped to a latent space via a variational autoencoder. Hence, we have

$$\mathbf{z}_t = \sqrt{\overline{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the Gaussian noise, $\mathbf{z}_0$ is the clean data in the latent space, $\mathbf{z}_t$ is the noisy latent feature at time step t, and $\overline{\alpha}_t$ is computed from a fixed variance schedule. During the backward process, a U-Net $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\cdot)$ [Ronneberger et al. 2015] is trained to restore information by predicting the noise at a time step t, with a loss of

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0, \boldsymbol{\epsilon}\sim\mathcal{N}, t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t, t)\|_2^2\right]. \tag{2}$$

To control the generation, a cross-attention layer is implemented in the U-Net to mix a control signal $\boldsymbol{\tau}_{\boldsymbol{\theta}}(\mathbf{y})$ with the intermediate layers of the U-Net, where $\mathbf{y}$ is commonly set to be a prompt embedding and $\boldsymbol{\tau}_{\boldsymbol{\theta}}(\cdot)$ is a trainable encoder. Therefore, a generative latent diffusion model can be trained with the following loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0, \mathbf{y}, \boldsymbol{\epsilon}\sim\mathcal{N}, t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t, t, \boldsymbol{\tau}_{\boldsymbol{\theta}}(\mathbf{y}))\|_2^2\right]. \tag{3}$$

Training such a diffusion model from scratch for every unique task is extremely resource-consuming. To adapt a pre-trained diffusion model to a new conditional task, one solution is to introduce a control branch via ControlNet [Zhang et al. 2023]. ControlNet preserves the quality and capabilities of the base model by locking its parameters. Furthermore, such a control branch makes a trainable copy of the U-Net encoding layers from the base model. The trainable copy is connected to the base model with zero convolution layers. The image generation process can then be optimized via

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0, \mathbf{y}, \boldsymbol{\epsilon}\sim\mathcal{N}, t, \mathbf{c}_\mathrm{I}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t, t, \boldsymbol{\tau}_{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{c}_\mathrm{I})\|_2^2\right], \tag{4}$$

where $\mathbf{c}_\mathrm{I}$ is the encoded control image.

3.2 Image Embedding

The original ControlNet adopts eight convolution layers to convert the image space condition into a feature conditioning vector. This provides a compact solution to capture high-level structure or style-related information, but loses some intrinsic information, as depicted in Figure 6. To address this issue, we propose a more effective conditional image encoder to extract dense features from the condition image without losing spatially-matched details. We draw inspiration from VQ-GAN [Esser et al. 2021] and employ a sequence of residual blocks [He et al. 2016] followed by multiple SwinV2 transformer layers [Liu et al. 2022] to generate feature-space conditioning vectors from the input RGB image. Specifically, our conditional image encoder consists of eight shared layers of residual blocks, $\mathbf{R}(\cdot)$, that map $\mathbf{I} \in \mathbb{R}^{W\times H\times 3}$ into $\mathbf{R}(\mathbf{I}) \in \mathbb{R}^{\frac{W}{8}\times\frac{H}{8}\times 320}$. These shared layers are followed by a domain-specific transformer $\mathbf{T}_*(\cdot)$ applied to $\mathbf{R}(\mathbf{I})$, where '$*$' is one of the intrinsic domains. We initialize the weights of the image encoder randomly during training. The output of this encoder $\mathbf{c}_\mathrm{I}^* = \mathbf{T}_*(\mathbf{R}(\mathbf{I}))$ is used as a condition for the ControlNet, as discussed next.

3.3 Joint Learning of Multiple Modalities

We train a ControlNet-like generative model to jointly learn three output modalities: albedo, shading, and surface normals. Similar to Zhang et al. [2023], we use a latent diffusion model [Rombach et al. 2022] as the base model and freeze the weights during training. We train the control branch by mixing the training data for different modalities with the corresponding prompts. Specifically, we use the name of the target modality, for example "Albedo", as the prompt text and encode it through the CLIP encoder [Radford et al. 2021]. This prompt is fixed and modality-specific. For a randomly sampled time step t, we follow the standard procedure to predict the noise added to the latent mapping, $\mathbf{z}_0^*$, of the corresponding ground-truth modality:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0^*, \mathbf{P}_*, \boldsymbol{\epsilon}\sim\mathcal{N}, t, \mathbf{c}_\mathrm{I}^*}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t^*, t, \mathbf{P}_*, \mathbf{c}_\mathrm{I}^*)\|_2^2\right], \tag{5}$$

where $\mathbf{c}_\mathrm{I}^*$ is the conditioning image feature vector, and $\mathbf{P}_*$ is the embedding of the corresponding prompt.

During the training of the U-Net $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\cdot)$, we enforce zero signal-to-noise ratio (SNR) when sampling the noise $\boldsymbol{\epsilon}$ [Du et al. 2023; Lin et al. 2024]. This improves the congruence between training and inference, and allows the model to generate samples that are more faithful to the original data distribution. V-prediction and V-loss strategies [Salimans and Ho 2022] are employed by the zero-SNR

scheduler [Lin et al. 2024] to ensure the model can learn a meaningful data distribution as the signal-to-noise ratio (SNR) approaches zero. Specifically, instead of predicting the noise, the velocity $\mathbf{v}$ is predicted via the diffusion network with

$$\mathbf{v}_t = \sqrt{\overline{\alpha}_t}\boldsymbol{\epsilon} - \sqrt{1-\overline{\alpha}_t}\mathbf{z}_0^*, \tag{6}$$

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0^*, \mathbf{P}_*, \boldsymbol{\epsilon}\sim\mathcal{N}, t, \mathbf{c}_\mathrm{I}^*}\left[\|\mathbf{v}_t - \tilde{\mathbf{v}}_{\boldsymbol{\theta}}(\mathbf{z}_t^*, t, \mathbf{P}_*, \mathbf{c}_\mathrm{I}^*)\|_2^2\right], \tag{7}$$

where $\tilde{\mathbf{v}}_{\boldsymbol{\theta}}(\cdot)$ is identical to $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\cdot)$ in terms of architecture, but trained for the velocity domain.

To accommodate the multiple intrinsic modalities in the same latent space of our diffusion model, we mix the different modalities (i.e., albedo, shading, surface normal) during training. In particular, we construct a single training batch by balancing the number of annotation-input pairs for each modality, which are obtained from various datasets (Section 4.1). Therefore, the residual blocks $\mathbf{R}(\cdot)$, the transformers $\mathbf{T}_A(\cdot)$, $\mathbf{T}_S(\cdot)$, and $\mathbf{T}_N(\cdot)$, and the ControlNet $\tilde{\mathbf{v}}_{\boldsymbol{\theta}}(\cdot)$ can be optimised jointly during the gradient back-propagation.

The shading values in a scene can span a wide range of values with a long-tailed distribution [Careaga and Aksoy 2023]. As our base latent diffusion model generates images in the range $[0, 1]$, we normalize the shading layer $\mathbf{S} \in [0, \infty)^3$ using $\mathbf{S}' = 1 - \frac{1}{1+\mathbf{S}}$. Note that the '1−' is added to a typical inverse shading formula [Careaga and Aksoy 2023] to keep the bright region in the original shading map still bright, i.e., a monotonic mapping.

3.4 Inference of the Intrinsic Modalities

The diffusion backward pass for each modality starts from an initial random 2D Gaussian noise and proceeds with the modality-specific conditioning image feature vectors $\mathbf{c}_\mathrm{I}^*$ and their corresponding text prompt embeddings $\mathbf{P}_*$. Our model adopts linear LDR images $\mathbf{I}$ as the image condition. Gamma correction is inverted through $\mathbf{I} = \mathbf{I}_{\mathrm{sRGB}}^{2.2}$ when the input in presented in the sRGB space. At each timestep, velocity is predicted, which is then converted to the intermediate latent noise maps following Salimans and Ho [2022]. Clean latent vectors are finally decoded to images using the base model's VQ-VAE [Esser et al. 2021]. To reduce the impact of randomness in initial noise sampling, we conduct diffusion processes for four initial random seeds, and average the output images of each modality.

4 EXPERIMENTS

4.1 Implementation Details

Our model is trained using 8 NVIDIA A100 GPUs with a batch size of 32 for 320K iterations, requiring approximately 65 hours. We train our ControlNet with the zero SNR strategy, on a base latent diffusion model also pre-trained with this strategy.

Training dataset. Our model is trained on a mix of datasets: (1) 52K synthesized images from InteriorVerse [Zhu et al. 2022] and 59K synthesized images from Hypersim [Roberts et al. 2021] with ground-truth albedo and surface normals, and ground-truth/computed HDR shading; (2) 9K real-world captured images from DIODE [Vasiljevic et al. 2019] with ground-truth surface normal maps. Images are resized to 384×384 pixels for training. Other implementation details are presented in Section A.1 in the *Supplement*.

4.2 Benchmark and Metrics

IIW benchmark. We compare our albedo estimation with previous works on the real-world IIW benchmark [Bell et al. 2014]. IIW is a scene-centric dataset, providing human judgements on the relative brightness of sparsely labelled pairs of albedo pixels, i.e., darker, similar or brighter. Following previous methods, we use the test split provided by Narihira et al. [2015], and evaluate methods using the standard WHDR (weighted human disagreement rate) metric. The WHDR metric measures the disagreement rate in the predicted intensity orderings compared to the labels. WHDR is a perceptual metric focusing solely on a partial property of the albedo, i.e., its consistency. However, this metric tends to favor flattened albedo [Wu et al. 2023] and cannot provide a comprehensive evaluation of albedo estimation (as discussed in Section A.2 in the *Supplement*).

SAW benchmark. We evaluate the shading estimation according to the average challenge precision (AP(c)) metric on the real-world SAW benchmark [Kovacs et al. 2017]. The SAW benchmark regards shading estimation as a binary classification problem: smooth or non-smooth, and provides sparse annotations. AP(c) calculates the average classification precision over 400 sampled smoothness thresholds. It was originally proposed by Li and Snavely [2018a] to focus on perceptual smoothness in challenging areas. During the evaluation, we use the official test split of this benchmark dataset.

ARAP benchmark. We follow Careaga and Aksoy [2023] to evaluate pixel-wise intrinsic image predictions on the synthetic ARAP benchmark [Bonneel et al. 2017] using dense ground truth. We use 123 test images from 37 high-quality Lambertian scenes, each with one or several illumination conditions. During evaluation, input images are resized with a maximum dimension of 1024 pixels. We study three scale-invariant quantitative metrics: mean-squared error (MSE), local mean-squared error (LMSE), and DSSIM, as proposed by Chen and Koltun [2013].

4.3 Comparisons

4.3.1 Baselines. We compare our approach to recent works including CGIntrinsics [Li and Snavely 2018a], CRefNet [Luo et al. 2023], NIID-Net [Luo et al. 2020], Ordinal Shading [Careaga and Aksoy 2023], PIE-Net [Das et al. 2022], and Zhu et al. [2022]. We evaluate these approaches using the code and trained models released by the authors on our benchmark datasets. Considering the characteristics of the different approaches, each model has been trained on datasets with dense annotations, sparse annotations, or multimodal annotations. Therefore, training on the same dataset is not the optimal setup for the individual methods.

4.3.2 Albedo estimation. We compare the albedo estimation performance on the real-world IIW benchmark [Bell et al. 2014] (as shown in Figure 7 and Table 1) and synthetic ARAP benchmark [Bonneel et al. 2017] (as shown in Table 2 and Figure 2 of the *Supplement*). NIID-Net [Luo et al. 2020] is not evaluated on the ARAP dataset, as the released checkpoint is trained with this dataset. We see that our predicted albedo clearly outperforms other alternatives with much better texture details and color consistency. Compared to other methods, our model shows better generalization to in-the-wild internet images, as shown in Figure 8. Quantitatively, our method achieves the best performance on two of the three albedo metrics in the ARAP benchmark, and achieves on-par performance with the

Table 1: Quantitative results on the IIW and SAW benchmarks. We evaluate albedo in the linear RGB space using the WHDR metric with both 10% and 20% equality thresholds. Various datasets are used to train the baselines, where IIW, SAW, and MI [2019] are real-world datasets, while NED [2018], GTA [2018], OR [2021b], Hypersim, and InteriorVerse are synthetic datasets. NYUv2 [2012] and DIODE are real-world datasets with only surface normal annotations. We highlight the top-2 scores in blue and the best in bold.

		WHDR ↓		
Method	Training data	10%	20%	AP(c) ↑
Zhu et al. [2022]	InteriorVerse	34.7	24.1	–
PIE-Net [2022]	NED+IIW	33.3	23.5	82.8
Li and Snavely [2018a]	CGI+IIW+SAW	23.9	16.5	97.9
Ordinal Shading [2023]	CGI+GTA+OR+Hypersim+MI	24.8	19.2	95.5
NIID-Net [2020]	CGI+NYUv2+DIODE	23.5	17.0	**98.4**
CRefNet [2023]	CGI+IIW	**12.8**	**10.8**	98.3
Ours	InteriorVerse+Hypersim+DIODE	17.9	13.3	98.3

Table 2: Quantitative results of albedo and shading estimation, and image reconstruction on the ARAP dataset. Our model achieves the best performance in five out of seven metrics. We highlight the best score in blue.

	Albedo			Shading			Reconst.
Method	MSE ↓	LMSE ↓	DSSIM ↓	MSE ↓	LMSE ↓	DSSIM ↓	MSE ↓
PIE-Net [2022]	0.042851	0.028001	0.159813	0.010657	0.005434	0.095594	0.000173
Ordinal Shading [2023]	0.036650	0.023871	0.155533	0.011423	0.005094	0.090179	0.000023
CRefNet [2023]	0.027791	0.016687	0.140501	0.010027	0.004971	0.091947	0.003869
Zhu et al. [2022]	0.027687	0.016514	0.136318	–	–	–	–
Ours	0.023577	0.014083	0.139976	0.007678	0.004153	0.080421	0.003054

Figure 4: The albedo predicted by Kocsis et al. does not match the input appearance according to the color and the content. Source: IIW test set.

existing state-of-the-art in terms of WHDR. Our method visually outperforms other methods significantly.

Compared with Careaga and Aksoy [2023], our approach does not assume that the chromaticity of the albedo is the same as that of the input image. Thus, ours allows for different chromaticity between the input and albedo, e.g., over-exposure (lamp in (b) of Figure 7), under-exposure (shadows in (a) of Figure 2 in *Supplement*), or complex intra-scene colorful reflections (bed in (b) of Figure 7). Our model generates images with better quality and inpaints the over-exposed area with reasonable content (c, d in Figure 7).

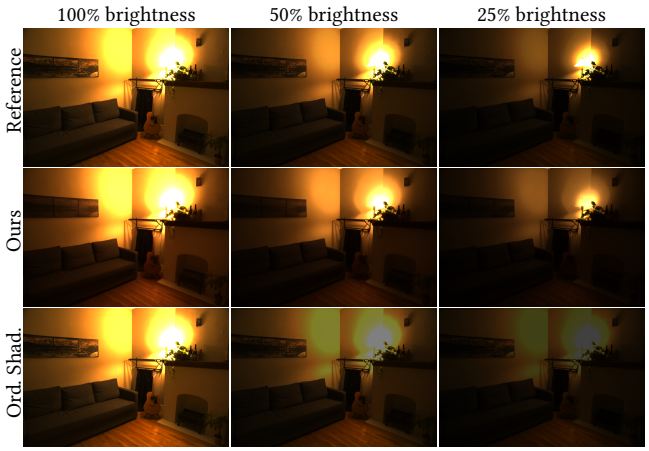

Figure 5: Comparison with Ordinal Shading [Careaga and Aksoy 2023] in brightness adjustment. Intrinsic images are estimated from the '100% brightness' LDR reference image (assumed to be in linear space). The estimation is then used to reconstruct the input with scaled shading intensity. Our model enables better high dynamic range adjustment due to the effective separation in the saturated region and recovery of the lost albedo information. Image source: Laval Indoor HDR Dataset [Garon et al. 2019].

CRefNet [Luo et al. 2023] adopts a strong albedo flattening prior, and trades off fine-grained albedo performance for improved consistency. Our model shows similar consistency but visually excels in fine-grained albedo reconstruction (coffee table in (a) in Figure 7 and top sample in Figure 1 of the *Supplement*). Note that the WHDR metric only considers paired samples, not reflecting the reproduction of details (as we discussed in Section A.2 in the *Supplement*), which explains the disparity between the qualitative results and the WHDR score. Furthermore, CRefNet has been trained using the training split of the IIW benchmark, whereas our method has not. Note that CRefNet visually underperforms compared to ours.

Finally, we provide a visual comparison to the concurrent work of Kocsis et al. [2024] in Figure 4, as it also leverages a diffusion model for intrinsic image decomposition. We observe that our albedo predictions are visually more consistent with the input image.

4.3.3 Shading estimation. In Table 1 and Table 2, we show that our method achieves competitive results on the SAW benchmark, and quantitatively the best on the three shading metrics on the ARAP benchmark. Since the AP(c) metric only measures the smoothness of the shading and cannot measure the correctness of the shading color, we also demonstrate visual comparisons in Figure 2 in the *Supplement*, where our shading results show a higher similarity to the ground truth. As shown in Figure 8, the estimation by the top-2 methods [Luo et al. 2020, 2023] on the SAW benchmark (with image resolutions less than 512) degrades when applied to high-resolution 1K test images. In particular, the shading estimated by Luo et al. [2020] relies on surface normal estimation, and significantly deteriorates when the quality of its surface normal estimation degrades at high resolutions. In contrast, our model exhibits consistent performance when applied to high-resolution inputs.

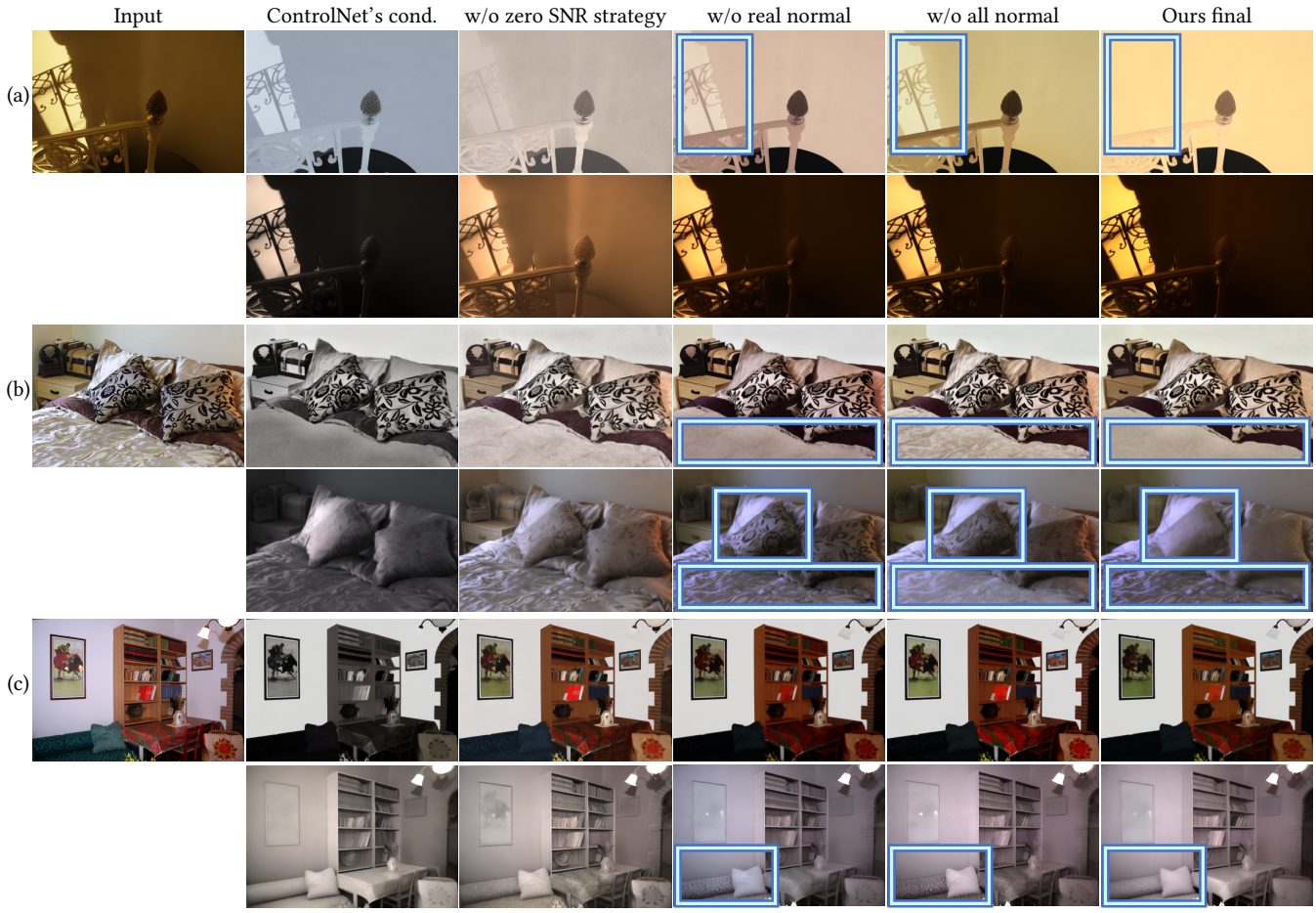

Figure 6: Visual comparison of our architecture design and training strategy ablations on the IIW test set. The albedo contrast in (b) has been enhanced for visual comparison. The ControlNet's image condition encoder causes significant color information loss, while color estimation is biased without the zero SNR strategy (a). As labelled, (a) more shadows are removed from the albedo; (b) input intensity variations are more accurately classified as caused by albedo variation or shading (shape) variation; (c) shading contains fewer texture residuals.

Table 3: Ablation studies on the architecture design and training strategy, evaluated on the ARAP (scale-invariant MSE, LMSE and DSSIM metrics), IIW (WHDR metric) and SAW (AP(c) metric) benchmarks. All the models are tested on images with the maximum dimension of 1024. 'A' indicates albedo estimation. 'S' indicates shading estimation. 'N' indicates surface normal estimation. For the training data, 'I' indicates InteriorVerse, 'H' indicates Hypersim, 'D' indicates DIODE. We highlight the top-2 scores in blue and the best in bold.

			Albedo			Shading			Reconstr.		
Method	Train data	Task	MSE ↓	LMSE ↓	DSSIM ↓	MSE ↓	LMSE ↓	DSSIM ↓	MSE ↓	WHDR ↓	AP(c) ↑
use ControlNet's image condition	I+H+D	A+S+N	0.029833	0.019146	0.148157	0.008580	0.004748	0.082044	0.004020	15.7	**98.6**
w/o zero SNR strategy	I+H+D	A+S+N	**0.023472**	**0.013130**	0.145799	0.009253	0.004935	0.095516	0.004992	14.5	98.0
w/o real surface normal data	I+H	A+S+N	0.025562	0.015425	**0.138207**	0.008182	0.004445	0.082764	0.002598	15.7	97.8
w/o all surface normal data	I+H	A+S	0.025311	0.015230	0.139936	0.007779	0.004175	0.081222	**0.002407**	14.6	98.2
Ours final	I+H+D	A+S+N	0.023577	0.014083	0.139976	**0.007678**	**0.004153**	**0.080421**	0.003054	**13.3**	98.3

As shown in Figure 7, Ordinal Shading and PIE-Net fail to accurately decode the underlying surface geometry, leading to non-smooth shading on planar areas (e.g., paintings in (b)). Our higher AP(c) score in Table 1 further supports this observation. Additionally, our approach achieves better global consistency. For example, the headboard's intensity in sample (b) should be similar to that of the wall behind and the bed below. This feature is not well-preserved by these two baseline methods.

It is worth noting that our method predicts a colorful HDR shading with RGB channels, which is essential in image editing tasks.

Ordinal Shading [2023] is also trained with the HDR shading data. However, as shown in Figure 5, our estimated shading for the oversaturated area has a wider range and can be used to adjust the brightness of the scene faithfully.

4.3.4 Surface normal estimation. Visual comparison with methods that jointly predict surface normals (Luo et al. [2020] and Li et al. [2020]) is shown in Figure 8. Our method achieves a significantly better performance while the baselines failed to generate meaningful outputs. Quantitative results are presented in Section A.4 in the *Supplement*.

4.4 Ablation Studies

4.4.1 Joint embedding of multiple modalities. One of our key insight is that jointly learning different modalities improves the overall quality. To verify this, we trained one model that does not predict surface normals and one model without using the real DIODE dataset. As shown in Figure 6 and Table 3, the realistic surface normal annotations from Vasiljevic et al. [2019] help the network perform better in albedo and shading estimation. Notably, surface normal estimation especially benefits sample (b), where frequent shading variations on the quilt challenge the classic "smooth shading" assumption. We believe our model gains a deeper understanding of shape variations from real-world DIODE data, enabling it to decide whether to predict smooth shading or not in each specific case.

4.4.2 Conditional image encoder. We also ablate the architecture of the image encoder $\mathbf{T}_*(\mathbf{R}(\cdot))$ by training a model variant adopting the ControlNet's original convolution-based encoder. As shown in Figure 6 and Table 3, our powerful image encoder better aligns the color between the input and the predictions.

4.4.3 Training with zero SNR. We further compare the models trained with and without the zero SNR noise scheduler. During training, we replace the base latent diffusion model with a version trained without the zero SNR but still in the velocity domain. Figure 6 shows that applying zero SNR provides a more effective strategy for aligning the color distribution of the intrinsic layers. Without this strategy, significant image reconstruction degradation occurs, as shown in Table 3.

4.5 Image Editing

Intrinsic layers are essential for various editing applications. We provide examples of relighting and retexturing in Figure 1. Implementation details are described in Section A.5 in the *Supplement*.

5 CONCLUSION AND LIMITATIONS

In this work, we presented an approach that leverages a pre-trained text-to-image foundation model to tackle the intrinsic image decomposition problem. We showed that the intrinsic information encoded in the foundational model can be effectively extracted through a novel conditioning mechanism that jointly predicts multiple intrinsic modalities. We evaluated our method thoroughly via qualitative and quantitative comparisons, and demonstrated editing applications. Given the nature of a data-driven approach, the performance of our model declines when it is applied to out-of-the-distribution scenarios, such as deeply folded fabrics, due to the domain gap between training and testing data. This will also affect the reconstruction of the input image from the intrinsic layers as the reconstruction error is not explicitly supervised during the model training.

ACKNOWLEDGMENTS

We would like to thank Li and Snavely [2018a], Li et al. [2020], Zhu et al. [2022], Das et al. [2022], and Careaga and Aksoy [2023] for publishing their source code. This work was supported by EPSRC CAMERA 2.0 (EP/T022523/1) and UKRI MyWorld Strength in Places Programme (SIPF00006/1).

REFERENCES

H. G. Barrow and J. M. Tenenbaum. 1978. Recovering intrinsic scene characteristics from images. *Computer Vision Systems* (1978).
Anil S Baslamisli, Thomas T Groenestege, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. 2018. Joint learning of intrinsic images and semantic segmentation. In *ECCV*. 286–302.
Ronen Basri and David W Jacobs. 2003. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 2 (2003), 218–233.
Sean Bell, Kavita Bala, and Noah Snavely. 2014. Intrinsic images in the wild. *ACM Transactions on Graphics* 33, 4 (2014), 159:1–12.
A.H. Bermano, R. Gal, Y. Alaluf, R. Mokady, Y. Nitzan, O. Tov, O. Patashnik, and D. Cohen-Or. 2022. State-of-the-Art in the Architecture, Methods and Applications of StyleGAN. *Computer Graphics Forum* 41, 2 (2022), 591–611. https://doi.org/10.1111/cgf.14503
Anand Bhattad, Daniel McKee, Derek Hoiem, and D.A. Forsyth. 2023. StyleGAN knows Normal, Depth, Albedo, and More. In *NeurIPS*.
Sai Bi, Xiaoguang Han, and Yizhou Yu. 2015. An L_1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM Transactions on Graphics* 34, 4 (2015), 78:1–12.
Nicolas Bonneel, Balazs Kovacs, Sylvain Paris, and Kavita Bala. 2017. Intrinsic decompositions for image editing. *Computer Graphics Forum* 36, 2 (2017), 593–609.
Adrien Bousseau, Sylvain Paris, and Frédo Durand. 2009. User-assisted intrinsic images. *ACM Transactions on Graphics* 28, 5 (2009), 130:1–10. https://doi.org/10.1145/1618452.1618476
Chris Careaga and Yağız Aksoy. 2023. Intrinsic Image Decomposition via Ordinal Shading. *ACM Trans. Graph.* 43, 1 (2023), 12:1–24. https://doi.org/10.1145/3630750
Qifeng Chen and Vladlen Koltun. 2013. A simple model for intrinsic image decomposition with depth cues. In *ICCV*. 241–248.
Changwoon Choi, Juhyeon Kim, and Young Min Kim. 2023. IBL-NeRF: Image-Based Lighting Formulation of Neural Radiance Fields. *Comput. Graph. Forum* (2023). https://doi.org/10.1111/cgf.14929
Partha Das, Maxime Gevers, Sezer Karaoglu, and Theo Gevers. 2023. IDTransformer: Transformer for Intrinsic Image Decomposition. In *ICCV Workshops*. 816–825.
Partha Das, Sezer Karaoglu, and Theo Gevers. 2022. PIE-Net: Photometric Invariant Edge Guided Network for Intrinsic Image Decomposition. In *CVPR*.
Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. 2023. Generative Models: What do they know? Do they know things? Let's find out!. In *NeurIPS*.
Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*. 12873–12883.
Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. 2018. Revisiting deep intrinsic image decompositions. In *CVPR*. 8944–8952.
David Forsyth and Jason Rock. 2022. Intrinsic Image Decomposition using Paradigms. *TPAMI* 44, 11 (2022), 7624–7637. https://doi.org/10.1109/TPAMI.2021.3119551
Elena Garces, Adolfo Muñoz, Jorge Lopez-Moreno, and Diego Gutierrez. 2012. Intrinsic Images by Clustering. *Computer Graphics Forum* 31, 4 (2012), 1415–1424.
Elena Garces, Carlos Rodriguez-Pardo, Dan Casas, and Jorge Lopez-Moreno. 2022. A Survey on Intrinsic Images: Delving Deep Into Lambert and Beyond. *International Journal of Computer Vision* 130 (2022), 836–868. https://doi.org/10.1007/s11263-021-01563-8
Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. 2019. Fast Spatially-Varying Indoor Lighting Estimation. In *CVPR*. 6908–6917.
Peter Vincent Gehler, Carsten Rother, Martin Kiefel, Lumin Zhang, and Bernhard Schölkopf. 2011. Recovering Intrinsic Images with a Global Sparsity Prior on Reflectance. In *NIPS*.
Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. 2009. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*. 2335–2342.
Mohammed Hachama, Bernard Ghanem, and Peter Wonka. 2015. Intrinsic scene decomposition from RGB-D images. In *ICCV*. 810–818.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. https://doi.org/10.1109/CVPR.2016.90

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *NeurIPS*. 6840–6851.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

Junqing Huang, Michael Ruzhansky, Qianying Zhang, and Haihui Wang. 2023. Intrinsic Image Transfer for Illumination Manipulation. *TPAMI* 45, 6 (2023), 7444–7456. https://doi.org/10.1109/TPAMI.2022.3224253

Yasamin Jafarian, Tuanfeng Y Wang, Duygu Ceylan, Jimei Yang, Nathan Carr, Yi Zhou, and Hyun Soo Park. 2023. Normal-guided Garment UV Prediction for Human Re-texturing. In *CVPR*.

Yeying Jin, Ruoteng Li, Wenhan Yang, and Robby T Tan. 2023. Estimating Reflectance Layer from A Single Image: Integrating Reflectance Guidance and Shadow/Specular Aware Learning. In *AAAI*.

Tero Karras, Samuli Laine, and Timo Aila. 2021. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 12 (dec 2021), 4217–4228. https://doi.org/10.1109/TPAMI.2020.2970919

Seungryong Kim, Kihong Park, Kwanghoon Sohn, and Stephen Lin. 2016. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *ECCV*. 143–159.

Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. 2024. Intrinsic Image Diffusion for Single-view Material Estimation. In *CVPR*.

Balazs Kovacs, Sean Bell, Noah Snavely, and Kavita Bala. 2017. Shading annotations in the wild. In *CVPR*. 6998–7007.

Philipp Krähenbühl. 2018. Free supervision from video games. In *CVPR*.

Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. 2018. DARN: a deep adversarial residual network for intrinsic image decomposition. In *WACV*. 1359–1367.

Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. 2021a. Semantic Segmentation with Generative Models: Semi-Supervised Learning and Strong Out-of-Domain Generalization. In *CVPR*.

Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2020. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In *CVPR*. 2475–2484.

Zhengqi Li and Noah Snavely. 2018a. CGIntrinsics: Better Intrinsic Image Decomposition Through Physically-Based Rendering. In *ECCV*.

Zhengqi Li and Noah Snavely. 2018b. Learning intrinsic image decomposition from watching the world. In *CVPR*. 9039–9048.

Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, Sai Bi, Hong-Xing Yu, Zexiang Xu, Kalyan Sunkavalli, Milos Hasan, Ravi Ramamoorthi, and Manmohan Chandraker. 2021b. OpenRooms: An Open Framework for Photorealistic Indoor Scene Datasets. In *CVPR*.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. 2024. Common Diffusion Noise Schedules and Sample Steps are Flawed. In *WACV*.

Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. 2020. Unsupervised learning for intrinsic image decomposition from a single image. In *CVPR*.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2022. Swin Transformer V2: Scaling Up Capacity and Resolution. In *CVPR*. 12009–12019.

Jundan Luo, Zhaoyang Huang, Yijin Li, Xiaowei Zhou, Guofeng Zhang, and Hujun Bao. 2020. NIID-Net: Adapting Surface Normal Knowledge for Intrinsic Image Decomposition in Indoor Scenes. *IEEE Transactions on Visualization and Computer Graphics* 26, 12 (2020), 3434–3445.

Jundan Luo, Nanxuan Zhao, Wenbin Li, and Christian Richardt. 2023. CRefNet: Learning Consistent Reflectance Estimation With a Decoder-Sharing Transformer. *IEEE Transactions on Visualization and Computer Graphics* (2023). https://doi.org/10.1109/TVCG.2023.3337870

Abhimitra Meka, Gereon Fox, Michael Zollhöfer, Christian Richardt, and Christian Theobalt. 2017. Live user-guided intrinsic video for static scenes. *IEEE Transactions on Visualization and Computer Graphics* 23, 11 (2017), 2447–2454.

Abhimitra Meka, Mohammad Shafiei, Michael Zollhöfer, Christian Richardt, and Christian Theobalt. 2021. Real-time Global Illumination Decomposition of Videos. *ACM Transactions on Graphics* 40, 3 (2021), 1–16.

Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. 2019. A Dataset of Multi-Illumination Images in the Wild. In *ICCV*. 4080–4089.

Takuya Narihira, Michael Maire, and Stella X Yu. 2015. Learning lightness from human judgement on relative reflectance. In *CVPR*. 2965–2973.

Ryan Po and Gordon Wetzstein. 2023. Compositional 3D Scene Generation using Locally Conditioned Diffusion. (2023). arXiv:2303.12218.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. 2021. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*. 10912–10922.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Tim Salimans and Jonathan Ho. 2022. Progressive Distillation for Fast Sampling of Diffusion Models. In *ICLR*.

Kripasindhu Sarkar, Marcel C. Buehler, Gengyan Li, Daoye Wang, Delio Vicini, Jérémy Riviere, Yinda Zhang, Sergio Orts-Escolano, Paulo Gotardo, Thabo Beeler, and Abhimitra Meka. 2023. LitNeRF: Intrinsic Radiance Decomposition for High-Quality View Synthesis and Relighting of Faces. In *SIGGRAPH Asia*. https://doi.org/10.1145/3610548.3618210

Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J. Fleet. 2023a. The Surprising Effectiveness of Diffusion Models for Optical Flow and Monocular Depth Estimation. In *NeurIPS*.

Saurabh Saxena, Junhwa Hur, Charles Herrmann, Deqing Sun, and David J. Fleet. 2023b. Zero-Shot Metric Depth with a Field-of-View Conditioned Diffusion Model. (2023). arXiv:2312.13252.

Viraj Shah, Svetlana Lazebnik, and Julien Philip. 2023. JoIN: Joint GANs Inversion for Intrinsic Image Decomposition. (2023). arXiv:2305.11321.

Jianbing Shen, Xiaoshan Yang, Yunde Jia, and Xuelong Li. 2011. Intrinsic images using optimization. In *CVPR*. 3481–3487.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from RGBD images. In *ECCV*. 746–760.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*.

Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. 2019. DIODE: A Dense Indoor and Outdoor DEpth Dataset. (2019). arXiv:1908.00463.

Zongji Wang, Yunfei Liu, and Feng Lu. 2023. Discriminative feature encoding for intrinsic image decomposition. *Computational Visual Media* 9 (2023), 597–618. https://doi.org/10.1007/s41095-022-0294-4

Chenglei Wu, Michael Zollhöfer, Matthias Nießner, Marc Stamminger, Shahram Izadi, and Christian Theobalt. 2014. Real-time shading-based refinement for consumer depth cameras. *ACM Transactions on Graphics* 33, 6 (2014), 200:1–10.

Jiaye Wu, Sanjoy Chowdhury, Hariharmano Shanmugaraja, David Jacobs, and Soumyadip Sengupta. 2023. Measured Albedo in the Wild: Filling the Gap in Intrinsics Evaluation. In *ICCP*.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Comput. Surv.* 56, 4, Article 105 (nov 2023), 39 pages. https://doi.org/10.1145/3626235

Weicai Ye, Shuo Chen, Chong Bao, Hujun Bao, Marc Pollefeys, Zhaopeng Cui, and Guofeng Zhang. 2023. IntrinsicNeRF: Learning Intrinsic Neural Radiance Fields for Editable Novel View Synthesis. In *ICCV*.

Lap-Fai Yu, Sai-Kit Yeung, Yu-Wing Tai, and Stephen Lin. 2013. Shading-based shape refinement of RGB-D images. In *CVPR*. 1415–1422.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*.

Qi Zhao, Ping Tan, Qiang Dai, Li Shen, Enhua Wu, and Stephen Lin. 2012. A closed-form solution to Retinex with nonlocal texture constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 7 (2012), 1437–1444.

Chengwei Zheng, Wenbin Lin, and Feng Xu. 2022. A Self-Occlusion Aware Lighting Model for Real-Time Dynamic Reconstruction. *IEEE Transactions on Visualization and Computer Graphics* (2022).

Hao Zhou, Xiang Yu, and David W Jacobs. 2019. GLoSH: Global-Local Spherical Harmonics for Intrinsic Image Decomposition. In *ICCV*. 7820–7829.

Tinghui Zhou, Philipp Krähenbühl, and Alexei A Efros. 2015. Learning data-driven reflectance priors for intrinsic image decomposition. In *ICCV*. 3469–3477.

Jingsen Zhu, Yuchi Huo, Qi Ye, Fujun Luan, Jifan Li, Dianbing Xi, Lisha Wang, Rui Tang, Wei Hua, Hujun Bao, and Rui Wang. 2023. I^2-SDF: Intrinsic Indoor Scene Reconstruction and Editing via Raytracing in Neural SDFs. In *CVPR*. https://doi.org/10.1109/CVPR52729.2023.01202

Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. 2022. Learning-based Inverse Rendering of Complex Indoor Scenes with Differentiable Monte Carlo Raytracing. In *Proceedings of SIGGRAPH Asia*. 6:1–8.

Michael Zollhöfer, Angela Dai, Matthias Innmann, Chenglei Wu, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2015. Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics* 34, 4 (2015), 96:1–14.

Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T Freeman. 2015. Learning ordinal relationships for mid-level vision. In *ICCV*.

Figure 7: Visual comparison of intrinsic image decomposition on the IIW test set [Bell et al. 2014]. For each sample, we show albedo (top) and shading (below) images. We show our surface normal predictions in the first column of each second row. Intrinsic image results are shown in the linear RGB space. Our model achieves the best perceptual performance. For example, our model effectively removes strong highlight shading from the estimated albedo (c and d) while preserving detailed wooden textures on the floor (c). Our shading is smooth and has minimal albedo residuals on the floor (a) and the paintings (b). Besides, our model captures the color of shading, which can benefit downstream image editing applications.

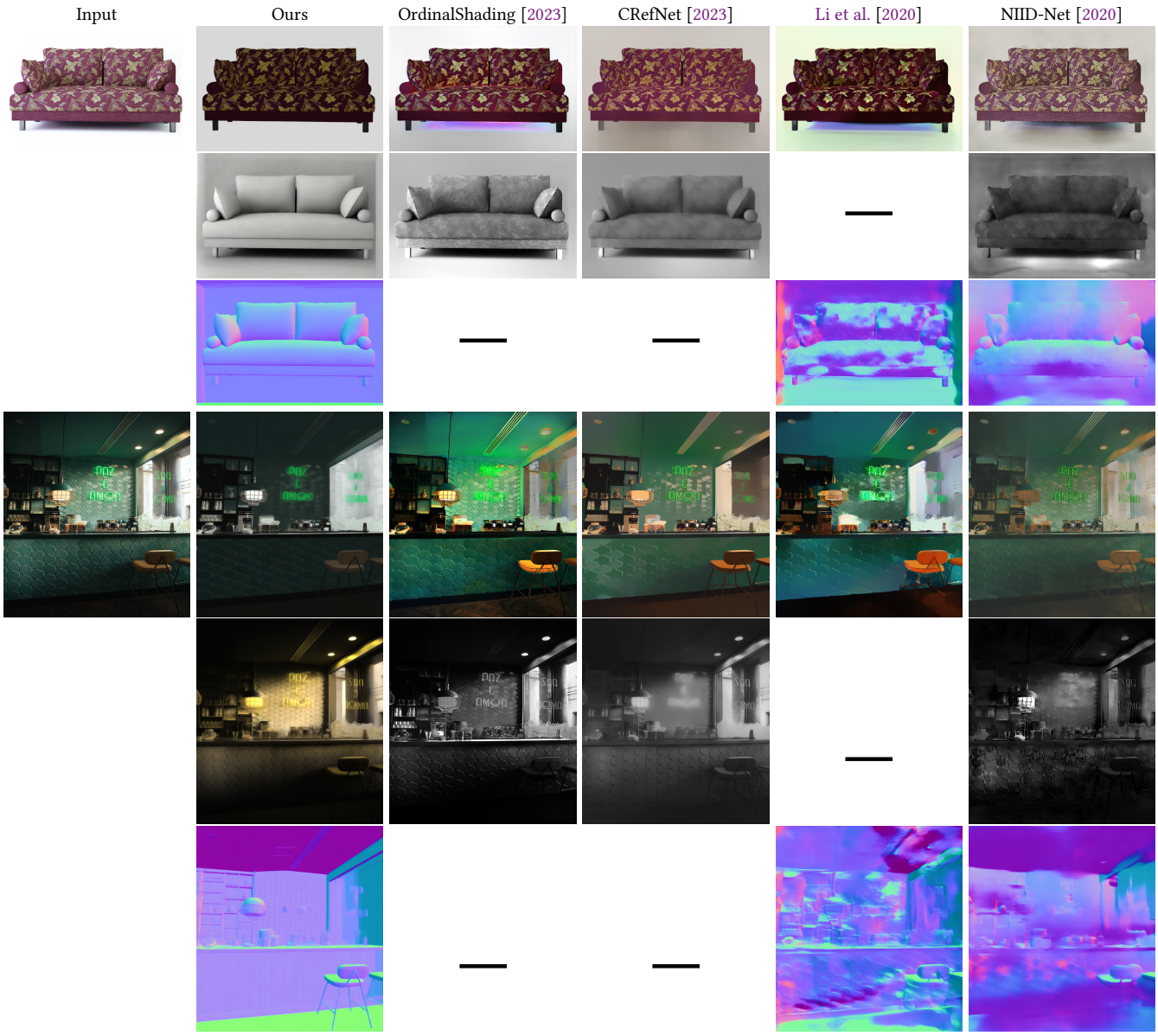

Figure 8: We evaluate our approach and the baselines with internet images at 1K resolution. For each sample, we show albedo (top), shading (middle) and surface normal (bottom) images. Intrinsic image results are shown in the linear RGB space. Our model is less sensitive to the input image resolution and consistently produces plausible intrinsic layer estimation. Image source: schanya, stock.adobe.com (top); Logan Stone, Unsplash (bottom).