RGB↔X: Image decomposition and synthesis using material- and lighting-aware diffusion models

Zheng Zeng
zhengzeng@ucsb.edu
Adobe Research
University of California, Santa Barbara
USA

Valentin Deschaintre
deschain@adobe.com
Adobe Research
United Kingdom

Iliyan Georgiev
igeorgiev@adobe.com
Adobe Research
United Kingdom

Yannick Hold-Geoffroy
holdgeof@adobe.com
Adobe Research
Canada

Yiwei Hu
yiwhu@adobe.com
Adobe Research
USA

Fujun Luan
fluan@adobe.com
Adobe Research
USA

Ling-Qi Yan
lingqi@cs.ucsb.edu
University of California, Santa Barbara
USA

Miloš Hašan
mihasan@adobe.com
Adobe Research
USA

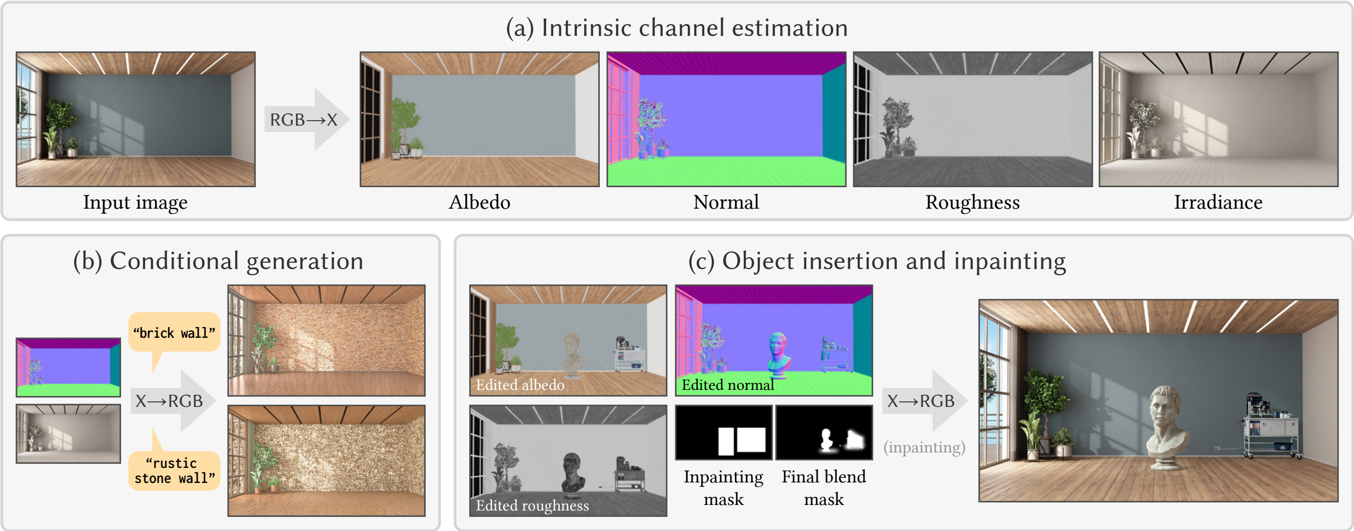


Figure 1: We present models for image decomposition into intrinsic channels (RGB→X) and image synthesis from such channels (X→RGB) in a unified conditional diffusion framework. (a) Our RGB→X model produces clean, plausible estimates of the intrinsic channels X. (b) New realistic images can be produced using our X→RGB model. Here we use a subset of the estimated channels, plus a text prompt. (c) We insert synthetic objects into the estimated channels and use an in-painting version of our X→RGB model, with appropriate masks, to synthesize a final composite image with matching lighting and shadows.

ABSTRACT

The three areas of realistic forward rendering, per-pixel inverse rendering, and generative image synthesis may seem like separate and unrelated sub-fields of graphics and vision. However, recent work has demonstrated improved estimation of per-pixel intrinsic channels (albedo, roughness, metallicity) based on a diffusion architecture; we call this the RGB→X problem. We further show that the reverse problem of synthesizing realistic images given intrinsic channels, X→RGB, can also be addressed in a diffusion framework. Focusing on the image domain of interior scenes, we introduce an improved diffusion model for RGB→X, which also estimates lighting, as well as the first diffusion X→RGB model capable of synthesizing realistic images from (full or partial) intrinsic channels. Our X→RGB model explores a middle ground between traditional rendering and generative models: We can specify only certain appearance properties that should be followed, and give freedom to the model to hallucinate a plausible version of the rest. This flexibility allows using a mix of heterogeneous training datasets that differ in the available channels. We use multiple existing datasets and extend them with our own synthetic and real data, resulting in a model capable of extracting scene properties better than previous work and of generating highly realistic images of interior scenes.

SIGGRAPH Conference Papers ’24, July 27-August 1, 2024, Denver, CO, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0525-0/24/07.
https://doi.org/10.1145/3641519.3657445

CCS CONCEPTS

• **Computing methodologies** → **Rendering**.

KEYWORDS

Diffusion models, intrinsic decomposition, realistic rendering

ACM Reference Format:
Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. 2024. RGB↔X: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27-August 1, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3641519.3657445

1 INTRODUCTION

Estimating geometric, shading, and lighting information from images has been long studied by the computer vision community, since classical work on intrinsic image decomposition. This problem is inherently difficult due to its under-constrained nature, including the ambiguity between illumination and materials [Grosse et al. 2009]. More recent work has focused on the related problem of per-pixel inverse rendering [Li et al. 2020; Zhu et al. 2022a]. This has produced physical material and lighting estimations, specifically diffuse albedos, specular roughness and metallicity, as well as various spatially varying representations of lighting. We refer to all of these information buffers as *intrinsic channels* and denote them using the symbol X, and the problem of estimating them as RGB⟶X.

On the other hand, computer graphics, and especially the subfield of physically based rendering, has long focused on the reverse task of turning detailed scene descriptions (comprising geometry, lighting, and materials) into realistic images. State-of-the-art rendering methods employ Monte Carlo light-transport simulation [Pharr and Humphreys 2004], commonly followed by a neural denoiser that encapsulates priors about plausible noise-free images. We refer to the problem of synthesizing an image from a given description as X⟶RGB.

A recent approach to producing highly realistic images, very different from traditional rendering, is based on generative models for image synthesis, especially based on large diffusion models [Ramesh et al. 2022; Rombach et al. 2022]. These models operate by iteratively denoising an image, pushing the neural-denoiser approach to the limit by starting from pure noise.

These three areas may seem unrelated, but we believe they should be studied in a unified way. We explore the connections between diffusion models, rendering, and intrinsic channel estimation, focusing on both material/light estimation and image synthesis conditioned on material/lighting, all in the same diffusion framework.

Recent work has demonstrated improved estimation of intrinsic channels based on a diffusion architecture. Kocsis et al. [2023] observe that further progress in this domain is likely to use generative modeling, due to the under-constrained and ambiguous nature of the problem. We follow this direction further. In addition to a new model for RGB⟶X which improves upon that of Kocsis et al. [2023], we also introduce a first X⟶RGB diffusion model which synthesizes realistic images from (full or partial) intrinsic channels. Much like RGB⟶X, the X⟶RGB problem requires a strong (ideally generative) prior to guide synthesis towards a plausible image, even with incomplete or overly simple intrinsic-channel information X.

Typical generative models are simple to use, but hard to precisely control. On the other hand, traditional rendering is precise but requires full scene specification, which is limiting. Our X⟶RGB model explores a middle ground where we specify only certain appearance properties that should be followed, and give freedom to the model to hallucinate a plausible version of the rest.

Our intrinsic channels X contain per-pixel albedo, normal vector, roughness, as well as *lighting* information which we represent as per-pixel irradiance on the scene surfaces. Furthermore, our X⟶RGB model is trained using channel dropout, which enables it to synthesize images using any subset of channels as input. This in turn makes it possible to use a mix of heterogeneous training datasets that differ in the available channels. We use multiple existing datasets and add our own synthetic and real data—a key advantage allowing us to expand training data beyond that of previous models. This paper makes the following contributions:

- An RGB⟶X model improving upon previous work [Kocsis et al. 2023] by using more training data from multiple heterogeneous datasets and adding support for lighting estimation;
- An X⟶RGB model capable of synthesizing realistic images from given intrinsic channels X, supporting partial information and optional text prompts. We combine existing datasets and add a new, high-quality interior scene dataset to achieve high realism.

In summary, we propose a unified diffusion-based framework that enables realistic image analysis (intrinsic channel estimation describing geometric, material, and lighting information) and synthesis (realistic rendering given the intrinsic channels), demonstrated in the domain of realistic indoor scene images; see Figure 1.

Our work is the first step towards unified frameworks for both image decomposition and synthesis. We believe it can bring benefits to a wide range of downstream editing tasks, including material editing, relighting, and realistic rendering from simple/under-specified scene definitions.

2 RELATED WORK

Generative models for images. Over the last decade, deep-learning-based image generation has rapidly progressed, notably with the advent of generative adversarial networks (GANs) [Goodfellow et al. 2014] and the subsequent body of research that improves both quality and stability [Gui et al. 2021; Karras et al. 2020; Pan et al. 2019]. However, the adversarial-based approach of GANs is prone to mode collapse, making them challenging to train. More recently, diffusion models have been shown to scale to training sets of hundreds of millions of images and produce extremely high-quality images [Ramesh et al. 2022; Rombach et al. 2022]. However, such models are costly to train, prompting research to fine-tune pre-trained models for various domains or conditioning [Hu et al. 2021; Sharma et al. 2023; Zhang et al. 2023], rather than training from scratch. We leverage the recent progress in this area to design our network architectures on top of Stable Diffusion v2.1 [Rombach et al. 2022], adding conditioning and dropout as a means for flexible input at test time.

Intrinsic decomposition. The problem of intrinsic image decomposition was defined almost five decades ago by Barrow et al. [1978] as a way to approximate an image I as a combination of diffuse reflectance (albedo), diffuse shading (irradiance), and optionally a specular term. Priors are necessary to estimate multiple values per pixel. Early priors include the retinex theory [Land and McCann 1971] which states that shading tends to have slower variation than reflectance. Pre-2009 methods are summarized by Grosse et al. [2009], while more recent methods are summarized by Garces et al. [2022]. We compare our albedo estimates to the most recent method of Careaga and Aksoy [2023].

Several recent works extend the traditional intrinsic decomposition to estimate more values per pixel, including specular roughness and/or metallicity, and lighting representations. Their training datasets focus on interior scenes. Li et al. [2020] are the first to use a large synthetic dataset of paired RGB renderings and decompositions to train a convolutional architecture for intrinsic channel estimation. The synthetic dataset used to train this method was later improved and released as *OpenRooms* [Li et al. 2021]. A further improvement was achieved by a switch from convolutional to vision transformer architectures [Zhu et al. 2022a]. More recently, Zhu et al. [2022b] introduce a new, more realistic synthetic interior dataset, and trained a convolutional architecture outperforming the method of Li et al. [2020], mostly due to the more realistic dataset.

A more recent alternative is to extract intrinsic images from pre-trained models such as StyleGAN [Karras et al. 2019] or pre-trained diffusion models [Bhattad et al. 2024; Du et al. 2023; Lee et al. 2023]. In this spirit, intrinsic image diffusion [Kocsis et al. 2023] proposes to fine-tune a general-purpose diffusion model to the per-pixel inverse rendering problem, going beyond previous methods by leveraging priors learned for image generation instead of predicting an average of the plausible solutions at each pixel. Their model is trained on InteriorVerse [Zhu et al. 2022b], a synthetic dataset of interior renderings. We further extend this work by training a similar RGB⟶X model with a different architecture on more data sources and additional intrinsic buffers. We further couple it with a new X⟶RGB model synthesizing realistic images from these buffers, effectively closing the loop back to RGB.

Normal estimation. Estimating per-pixel normal is related to intrinsic decomposition as it estimates 3D information for each pixel which is highly relevant to shading. However, this problem is typically studied in isolation from intrinsic images and has recently received limited attention compared to depth estimation. To demonstrate the competitiveness of our method, we consider an internal method, PVT-normal, based on Pyramid Vision Transformer [Wang et al. 2022] and trained on datasets similar to MiDaS [Birkl et al. 2023; Ranftl et al. 2022] to estimate normals. In our tests, PVT-normal outperforms the currently available state-of-the-art normal estimation methods. This model is not specific to interior scenes and is trained on a diverse dataset.

Neural image synthesis from decompositions. Several previous works have explored problems similar to our X⟶RGB problem. Deep Shading [Nalbach et al. 2017] solves the problem of learning screen-space shading effects (e.g., ambient occlusion, image-based lighting, subsurface scattering) using a CNN-based architecture learned on synthetic data, resulting in fast rendering, competitive or better than hand-tuned screen-space shaders. Deep Illumination [Thomas and Forbes 2018] is an approach based on a conditional GAN learned per scene, efficiently predicting global illumination given screen-space intrinsic buffers, while direct illumination is computed analytically. Zhu et al. [2022b] introduce a screen-space ray-tracing approach to synthesize images from intrinsic channels. In contrast, our approach jointly considers image decomposition and synthesis, does not require any ray tracing, and its models are general across the interior scene domain.

Table 1: We combine four heterogeneous datasets (ours in bold), each providing a subset of the channels we need for training. For each dataset we mark channels as available (✓), unavailable (✗), or available but not fully reliable (✓). We also include representative images from the datasets. ImageDecomp is an RGB-only dataset for which we estimated the intrinsic channels using our RGB⟶X model.

Dataset	Size	Albedo	Normal	Roughness	Metallic.	Irrad.
InteriorVerse	50,097	✓	✓	✓	✓	✗
Hypersim	73,819	✓	✓	✗	✗	✓
Evermotion	17,000	✓	✓	✓	✓	✗
ImageDecomp	50,000	✓	✓	✓	✓	✓

Relighting. Single-image scene relighting methods have been proposed using both explicit [Griffiths et al. 2022; Pandey et al. 2021; Yu et al. 2020] and implicit [Rudnev et al. 2022; Wang et al. 2023] representations. These works are limited to simple lighting: a single directional light source or low-order spherical harmonics. Closer to our work, Li et al. [2022] build a per-pixel inverse rendering method to relight interior scenes from a single image. Furthermore, they introduce a hybrid neural and classical rendering system that synthesizes relit images given intrinsic channels and lighting information, similar to our X⟶RGB. While we believe our framework can be part of a toolbox for relighting, we do not specifically focus on solving the relighting problem, which poses challenges beyond our scope.

3 INTRINSIC CHANNELS AND DATASETS

In this section, we discuss the intrinsic channels X used in our models, and the datasets with paired RGB images and intrinsic channels that we used or prepared ourselves.

3.1 Intrinsic channels

In our RGB⟶X and X⟶RGB models, we use the following channels:

- Normal vector $\mathbf{n} \in \mathbb{R}^{H \times W \times 3}$ specifying geometric information in camera space;
- Albedo $\mathbf{a} \in \mathbb{R}^{H \times W \times 3}$, also commonly referred to as base color, which specifies the diffuse albedo for dielectric opaque surfaces and specular albedo for metallic surfaces;

- Roughness $\mathbf{r} \in \mathbb{R}^{H \times W}$, typically understood as the square root of the parameter α in GGX or Beckmann microfacet distributions [Walter et al. 2007]. High roughness means more matte materials while low roughness means shinier;
- Metallicity $\mathbf{m} \in \mathbb{R}^{H \times W}$, typically defined as a linear blend weight interpolating between treating the surface as dielectric and metallic; and
- Diffuse irradiance $\mathbf{E} \in \mathbb{R}^{H \times W \times 3}$, which serves as a lighting representation. It represents the amount of light reaching a surface point integrated over the upper cosine-weighted hemisphere.

We also contemplated adding a per-pixel depth channel, but eventually found it unnecessary, as depth can be estimated from normals, and normals typically contain more information about high-frequency local variations.

Unlike a material system in a traditional rendering framework, the above properties are fairly imprecise. For example, they cannot represent glass. Instead, we treat glass as having zero roughness and metallicity. This usually does not pose problems: the model infers from context that an object is a window or a glass cabinet, and plausibly inpaints objects or illumination behind the glass.

All intrinsic channels in our datasets have the same resolution as the corresponding RGB images, and are estimated at full resolution by RGB→X. However, it is sometimes beneficial to condition X→RGB on downsampled channels, as discussed in Section 5.

3.2 Datasets

To train our models, we ideally desire a large, high-quality image dataset, containing paired information for all of the channels we require: normal $\mathbf{n}$, albedo $\mathbf{a}$, roughness $\mathbf{r}$, metallicity $\mathbf{m}$, diffuse irradiance $\mathbf{E}$, the corresponding RGB image $\mathbf{I}$ (ideally a real photograph or at least a very realistic render), and a text caption describing the image. However, no existing dataset satisfies these requirements, and we instead piece together datasets with partial information and construct new datasets to fill the gaps. Table 1 summarizes the size and channel availability of the datasets we use.

InteriorVerse [Zhu et al. 2022b] is a synthetic indoor scene dataset, containing over 50,000 rendered images with $\mathbf{n}$, $\mathbf{a}$, $\mathbf{r}$, and $\mathbf{m}$ channels in addition the rendered images $\mathbf{I}$. There are a few issues with this dataset. First, the rendered images contain noise; this does not pose problems for RGB→X estimation, but the X→RGB synthesis model learns to reproduce the undesirable noise. We resolve this by applying an off-the-shelf denoiser (NVIDIA OptiX denoiser [NVIDIA 2020]). Furthermore, we found that roughness and metallicity values are often dubious, and decided not use them for this dataset. The dataset also has a synthetic style, which the X→RGB model would learn to imitate if trained exclusively on it. The small variety of objects and materials causes some biases, e.g., green albedo has strong correlation with plants, so a green-albedo wall would be synthesized with a leafy texture if trained solely on InteriorVerse.

Hypersim [Roberts et al. 2021] is another synthetic photorealistic dataset comprising over 70,000 rendered images, with $\mathbf{n}$, $\mathbf{a}$, and most importantly $\mathbf{E}$ data available. This dataset does not include other material information like roughness and metallicity, and sometimes bakes specular shading into the albedo. Fortunately, this is not common enough to preclude us from using the albedo data. While Hypersim expands the scene appearance variety over InteriorVerse, it is still not sufficient for highly realistic synthesis.

We complete these with two of our own datasets. The first is Evermotion, a synthetic dataset generated similarly to InteriorVerse by rendering synthetic scenes created by artists, randomly placing cameras along pre-recorded camera paths, and rendering 17,000 images of 85 indoor scenes. The main benefit of Evermotion is that it provides us with roughness $\mathbf{r}$ and metallicity $\mathbf{m}$, for which this dataset is currently the only reliable source.

To further enhance the training data and help our X→RGB model synthesize realistic images, we use 50,000 high-quality commercial interior scene images. These images come from photographs or high-quality renderings, with no additional channels available. We therefore estimate normals, albedo, roughness, metallicity, and diffuse irradiance using our RGB→X model. The combination of images and estimated channels form our ImageDecomp dataset.

To better preserve the existing text-understanding abilities of the base diffusion model during fine-tuning for X→RGB, we precompute image captions for all images in all of the above datasets, using the BLIP-2 model [Li et al. 2023].

4 THE RGB→X MODEL

In this section, we describe our RGB→X model to estimate the intrinsic channels X from an input RGB image $\mathbf{I}$. The output contains all channels discussed in Section 3.1. Similarly to Kocsis et al. [2023], we fine-tune a pre-trained text-to-image latent diffusion model, Stable Diffusion 2.1 [Rombach et al. 2022]. Figure 2 shows a high-level overview of our model.

Like Stable Diffusion, our RGB→X model operates on a latent space with a pre-trained encoder $\mathcal{E}$ and decoder $\mathcal{D}$. During inference, it takes $\mathcal{E}(\mathbf{I})$ as input condition and iteratively denoises a Gaussian-noise latent image $\mathbf{z}_T^{\mathrm{X}}$ to produce the target latent image $\mathbf{z}_0^{\mathrm{X}}$ encoding the intrinsic channels X. During training, we optimize the following loss function L_θ with v-prediction [Salimans and Ho 2022] as we find v-prediction to give better results than noise ϵ prediction:

$$\mathbf{v}_t^{\mathrm{RGB}\to\mathrm{X}} = \sqrt{\bar{\alpha}_t}\epsilon - \sqrt{1-\bar{\alpha}_t}\mathbf{z}_0^{\mathrm{X}} \,, \tag{1}$$

$$L_\theta = \left\| \mathbf{v}_t^{\mathrm{RGB}\to\mathrm{X}} - \hat{\mathbf{v}}_\theta^{\mathrm{RGB}\to\mathrm{X}} \left(t, \mathbf{z}_t^{\mathrm{X}} \mathcal{E}\left(\mathbf{I}\right), \tau\left(\mathbf{p}^{\mathrm{X}}\right)\right) \right\|_2^2 \,. \tag{2}$$

Here t is the noise amount ("time step") drawn uniformly during training, $\epsilon \sim \mathcal{N}(0,1)$, $\bar{\alpha}_t$ is a scalar function of t, $\hat{\mathbf{v}}_\theta^{\mathrm{RGB}\to\mathrm{X}}$ is our RGB→X diffusion model with parameters θ, $\mathbf{z}_0^{\mathrm{X}}$ is the target latent, $\mathbf{z}_t^{\mathrm{X}}$ is the noisy latent after adding noise ϵ at time step t to $\mathbf{z}_0^{\mathrm{X}}$. Further, $\mathbf{p}^{\mathrm{X}}$ is the text prompt computed for $\mathbf{I}$, and τ is the CLIP text encoder [Radford et al. 2021] that encodes the prompt into a text-embedding vector. This CLIP embedding is used as a specialized context for cross-attention layers of the model.

Image and intrinsic channel encoding. As our model operates in latent space, we encode the input image as $\mathcal{E}\left(\mathbf{I}\right)$ and concatenate it to the noisy latent $\mathbf{z}_t^{\mathrm{X}}$ as input to our $\hat{\mathbf{v}}_\theta^{\mathrm{RGB}\to\mathrm{X}}$ model. We use the frozen encoder $\mathcal{E}$ from the original Stable Diffusion model, which we also found to work well for all our intrinsic images encoding.

Handling multiple output channels. The output of the original Stable Diffusion model is a 4-channel latent image which can be decoded into a single RGB image. As we aim to produce additional

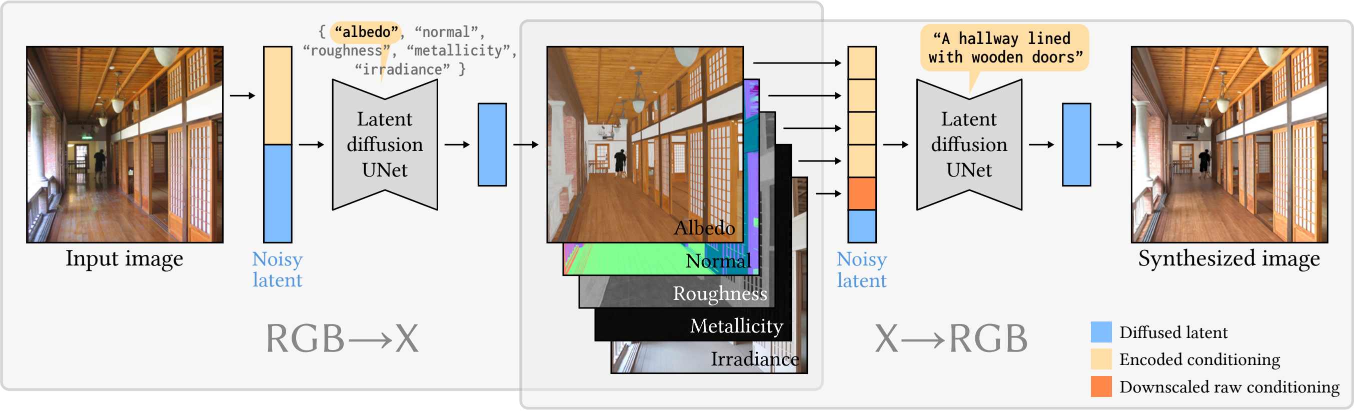


Figure 2: High-level overview of our two diffusion models. Left: The RGB⟶X model takes the input image, encoded into latent space by the pre-trained encoder, concatenated with the diffusion latent. We repurpose the text prompt to a switch choosing the desired output channel; this allows training with datasets containing any subset of the supported channels. Right: The X⟶RGB model concatenates the input intrinsic channels, again encoded by the pre-trained encoder, with the diffusion latent. One exception is the irradiance (lighting) channel, which is downsampled to latent resolution rather than passed through the encoder. This model can accept usual text prompts. All input conditions to X⟶RGB are optional.

output channels (albedo **a**, normal **n**, roughness **r**, metallicity **m**, and lighting **E**) we may expect that a larger latent vector may help better encode the information as done in previous work [Kocsis et al. 2023]. However, we find that extending the number of latent channels of the original model leads to lower-quality results. Indeed, adding more latent channels to the operating latent space of a diffusion model forces us to re-train both input and output convolutional layers from scratch. In a way, the model is suddenly "shocked" into a new domain, making the training more challenging.

We train our model with various datasets to increase variety, as described in Section 3.2, but this comes with the additional issue of heterogeneous intrinsic channels, which is challenging for our approach that stacks all intrinsic channels into a larger latent. A straightforward approach would be to only include the loss for available maps in each training iteration. We however found this approach to perform poorly.

Our solution is to produce a single intrinsic channel at a time and repurpose the input text prompt (which does not serve any other purpose in our RGB⟶X task) as a "switch" to control the diffusion-model output. Previous work [Brooks et al. 2023; Sharma et al. 2023] shows that it is possible to use specially designed prompts as instructions to control a single diffusion model to perform different tasks explicitly. Inspired by this design, we use five fixed prompts acting as switches. More specifically, one unit of data is collated as $\{\mathbf{g}, \mathbf{p}^{\mathrm{X}}, \mathbf{I}\}$ where $\mathbf{g} \in \{\mathbf{n}, \mathbf{a}, \mathbf{r}, \mathbf{m}, \mathbf{E}\}$ and $\mathbf{p}^{\mathrm{X}} \in$ { "normal", "albedo","roughness", "metallicity", "irradiance"} is set accordingly. We find that this approach performs similarly to fine-tuning separate models for each output modality in $\{\mathbf{n}, \mathbf{a}, \mathbf{r}, \mathbf{m}, \mathbf{E}\}$ while fine-tuning and storing only a single network's weights.

5 THE X⟶RGB MODEL

We now describe our X⟶RGB model, performing realistic RGB image synthesis from intrinsic channels X, illustrated in Figure 2. Much like for RGB⟶X, we fine-tune a diffusion model starting from Stable Diffusion 2.1 with several different considerations.

In the X⟶RGB case, we define the target latent variable as $\mathbf{z}_0^{RGB} = \mathcal{E}(\mathbf{I})$, directly encoding the image **I**. We provide the input condition X through concatenation of the encoded input intrinsic channels, adjusted to take various dataset properties in consideration as described below. When using all intrinsic images, the input latent vector is defined as

$$\mathbf{z}_t^{\mathrm{X}} = (\mathcal{E}(\mathbf{n}), \mathcal{E}(\mathbf{a}), \mathcal{E}(\mathbf{r}), \mathcal{E}(\mathbf{m}), \mathcal{E}(\mathbf{E})). \tag{3}$$

We train our X⟶RGB model by minimizing the loss function L'_θ:

$$\mathbf{v}_t^{\mathrm{X}\to\mathrm{RGB}} = \sqrt{\bar{\alpha}_t}\epsilon - \sqrt{1-\bar{\alpha}_t}\mathbf{z}_0^{RGB} , \tag{4}$$

$$L'_\theta = \left\| \mathbf{v}_t^{\mathrm{X}\to\mathrm{RGB}} - \hat{\mathbf{v}}_\theta^{\mathrm{X}\to\mathrm{RGB}}\left(t, \mathbf{z}_t^{RGB}, \mathbf{z}_t^{\mathrm{X}}, \tau(\mathbf{p})\right) \right\|_2^2 . \tag{5}$$

Here, $\hat{\mathbf{v}}_\theta^{\mathrm{X}\to\mathrm{RGB}}$ is our X⟶RGB diffusion model with parameters θ. We concatenate the noisy latent $\mathbf{z}_t^{RGB}$ and the conditioning latent $\mathbf{z}_t^{\mathrm{X}}$ together before feeding it into $\hat{\mathbf{v}}_\theta^{\mathrm{X}\to\mathrm{RGB}}$. The CLIP text embedding $\tau(\mathbf{p})$ is used as the context for cross-attention layers of the model. For X⟶RGB the text embedding is used as an additional control as is usual in Diffusion Models.

While our RGB⟶X model required a solution to output multiple modalities, the X⟶RGB model only requires changing the input layers to handle the additional conditional latent channels. Indeed, as in the original Stable Diffusion, the output remains a single RGB image. During training, only the newly added weights of the input convolutional layer need to be trained from scratch to handle the additional conditions, which does not "shock" the model out of its normal denoising ability for $\mathbf{z}_t^{RGB}$.

Handling heterogeneous data. Still, the problem of different intrinsic data channels missing from different datasets remains. To resolve this issue, we follow the observations made in previous works [Ho and Salimans 2022; Huang et al. 2023] which propose to jointly train a conditional and unconditional diffusion model through condition channel dropout to improve sample quality and enable image generation with any subset of conditions. We therefore use an intrinsic channel drop-out strategy; with it, our conditioning latent $\mathbf{z}_t^{\mathsf{X}}$ can be rewritten as:

$$\mathcal{P}(x) \in \{\mathcal{E}(x), 0\} \tag{6}$$

$$\mathbf{z}_t^{\mathsf{X}} = (\mathcal{P}(\mathbf{n}), \mathcal{P}(\mathbf{a}), \mathcal{P}(\mathbf{r}), \mathcal{P}(\mathbf{m}), \mathcal{P}(\mathbf{E})). \tag{7}$$

This approach lets us handle heterogeneous datasets during training, and choose which inputs to provide at inference; for example, providing no albedo or no lighting will result in the model generating plausible images, using its prior to compensate for the missing information (Figure 6).

Low-resolution lighting. Our RGB→X model succeeds in estimating highly detailed lighting in the form of a diffuse irradiance image **E**, closely following high-resolution geometry and normals. While this could be beneficial for some applications, using these detailed lighting buffers for X→RGB presents an issue if we want to actually *edit* the detailed normals and control the lighting using a coarser interpretation of **E**. In other words, we would like to provide the lighting as a "hint" to the X→RGB model, rather than a precise per-pixel control. Instead of encoding the full resolution lighting **E** into the latent space as for other conditions, we simply downsample it into the same resolution as the latent. By doing so, we provide the X→RGB model with a coarser hint of lighting without pixel detail, while still achieving adherence to the overall lighting condition. This is important, e.g., when editing the normals in Figure 7.

Fine-tuning for inpainting. To enable local editing applications (shown in Figures 1 and 7) we fine-tune our X→RGB model to support inpainting by simply adding a masked image and mask channels to the model input. We downsample the mask to the latent-space resolution and concatenate it to the conditioning latent $\mathbf{z}_t^{\mathsf{X}}$.

6 RESULTS

Note about picking results from a generative model. Applying generative models to the RGB→X and X→RGB problems means that the output is not unique but sampled from a distribution. While we could evaluate a number of samples and take their mean [Kocsis et al. 2023], we do not recommend this approach, as it can blur details that have been reasonably estimated within each sample. Instead, we pick a single sample to display in the paper, and provide more samples in the supplementary materials. Albedo, lighting, and normal samples are typically usable, while more attention is required for roughness and metallicity due to the lack of reliable training data and the inherent ambiguity of these properties.

6.1 RGB→X on synthetic and real inputs

Figures 3 and 4 show our results on intrinsic channel estimation for synthetic and real examples. None of the synthetic input examples were part of the training data. Please see the supplementary materials for many more results.

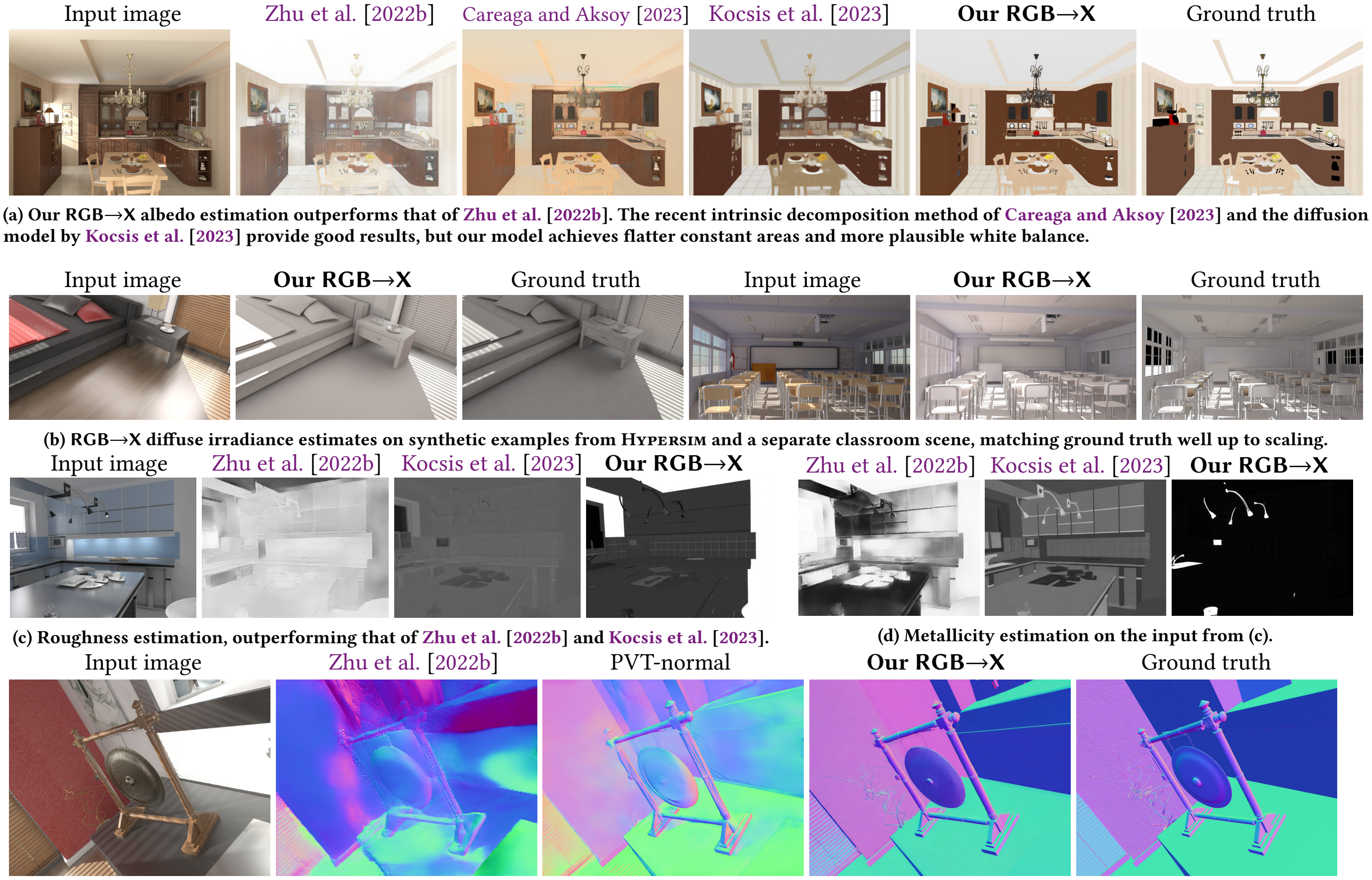
Albedo. We compare albedo estimation to previous work in Figure 3(a) for synthetic and Figure 4(a, b) for real inputs. Generally, we find that our model is best at removing reflections, highlights, shadows, and color cast from the inputs, while providing the flattest estimates for albedo regions that should indeed be constant. The method of Zhu et al. [2022b] performs worse on both synthetic and real inputs, hinting at the limitations of non-generative models, nor have designs incorporating special knowledge about the albedo estimation problem. The recent intrinsic decomposition method of Careaga and Aksoy [2023] provides good results, but our model achieves flatter constant areas and a more plausible white balance. Although they also show impressive results, the same is true for the diffusion model by Kocsis et al. [2023]. For example, on the bedroom photo in Figure 4(a, top row), our model is the only one correctly predicting that all bed-linen pixels should have identical white albedo. The challenging real image in Figure 4(b) also results in a very clean albedo estimate that outperforms other methods, though our model removes some wear from the wooden floor, possibly due to training on synthetic materials without wear.

Diffuse irradiance (lighting). In Figure 3(b), we see that our model produces diffuse irradiance estimates that closely match ground truth on synthetic data, even on inputs with intricate shadow patterns, and with very little to no leaking of material properties into the estimation. The color in the irradiance is also plausibly shifted away from pure white under colored lighting. Our estimates are also realistic and plausible on real inputs, such as in Figure 4(b). Careaga and Aksoy [2023] do not provide irradiance directly, so we divide the original image by their predicted albedo and use the resulting approximate irradiance as a baseline.

Metallicity and roughness. As shown in Figure 3(c,d) and Figure 4(c,d), our RGB→X model generates much more plausible roughness and metallicity for a given input image than the previous publicly available state of the art [Kocsis et al. 2023; Zhu et al. 2022b]. These material properties are challenging to recover accurately, for two reasons. First, the amount of reliable training data for them is the lowest. Second, they only impact surface reflection significantly if lit by appropriate high-frequency illumination; otherwise the model has to revert to prior knowledge, estimating what the object could be and whether such objects tend to be rough or metallic. These issues translate to a higher sampling variance of our model, and a lower yield of "good" samples. We show this variability in our estimations in the supplemental materials.

Normals. In synthetic tests (Figure 3(e)) as well as real ones (Figure 4(e)), we show that our model estimates normals plausibly, including high-frequency geometry, while correctly predicting flat normals for flat surfaces even if they have texture or high-frequency lighting. Our results outperform those of Zhu et al. [2022b] and slightly improve on the state-of-the-art PVT-normal. While we observe that our model normal estimation generalizes reasonably well (see more examples in supplementary materials), we do not claim general improvement in this space, as PVT-normal is specifically designed to work well across general images. We provide this comparison for the sake of completeness.

Quantitative comparisons. For albedo, normal, roughness and metallicity estimation, we compare to the corresponding previous methods in Table 2. We find that our RGB→X has the best PSNR and LPIPS values on all channels, with the exception of irradiance for which we do not have existing methods to compare to.



(a) Our RGB→X albedo estimation outperforms that of Zhu et al. [2022b]. The recent intrinsic decomposition method of Careaga and Aksoy [2023] and the diffusion model by Kocsis et al. [2023] provide good results, but our model achieves flatter constant areas and more plausible white balance.

(b) RGB→X diffuse irradiance estimates on synthetic examples from Hypersim and a separate classroom scene, matching ground truth well up to scaling.

(c) Roughness estimation, outperforming that of Zhu et al. [2022b] and Kocsis et al. [2023].

(d) Metallicity estimation on the input from (c).

(e) RGB→X normal estimation, outperforming Zhu et al. [2022b], as well as a SOTA model based on vision transformers (trained on much more diverse data).

Figure 3: Synthetic data comparison of our RGB→X model against previous methods [Careaga and Aksoy 2023; Zhu et al. 2022b] and a known ground truth. All input images and ground truths are from Hypersim, except for the classroom scene (c).

6.2 X→RGB model results

Comparison to path tracing reference. In Figure 5, we validate that our X→RGB model produces results closely matching traditional Monte Carlo path tracing, as long as the input channels X are not far from the training distribution of synthetic interiors. Here, we use a common synthetic kitchen scene, not part of our training data. We use all intrinsic channels (shown on the left) and feed them into our model, along with a text prompt. The result matches the path-traced reference well in terms of material appearance and global illumination. Differences can also be noted: for example, the stove has a dark metallic material in the input channels, which is rare in the training data. Our model generates a brighter aluminum material, matching the metallicity instead of the albedo channel.

Subsets of input channels and text prompts. Figure 6 demonstrates the ability of our X→RGB model to generate plausible images by specifying only a subset of the appearance properties as input. Furthermore, text prompts can be used for additional control. Here, we control the lighting (a) or object colors (b). Generally, text control works well only when there are only few objects (e.g., one sofa and a few pillows). It is hard to control the color of a specific object by text, but this issue is a common challenge for all diffusion models.

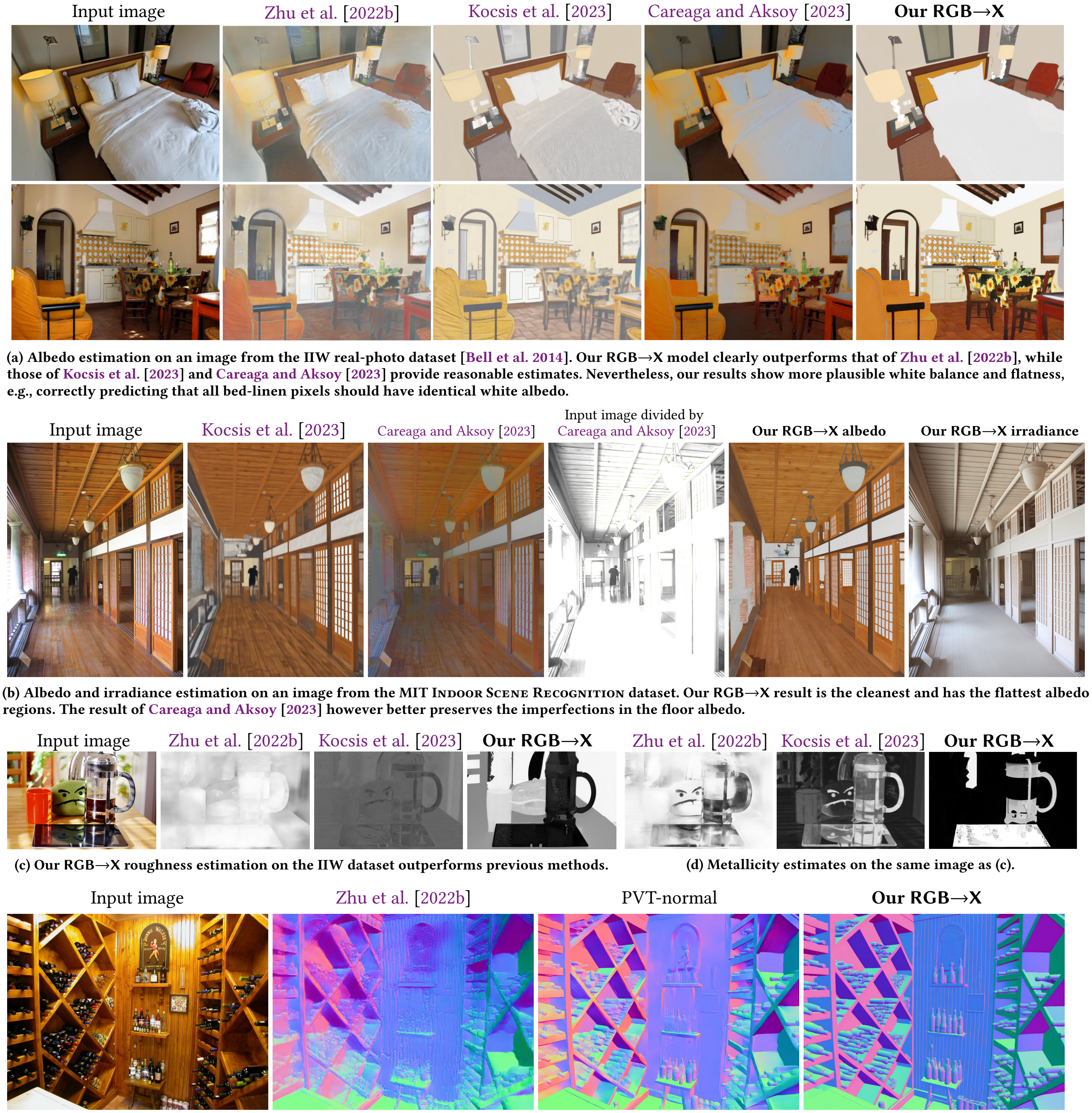
Table 2: Numerical evaluation of RGB→X against existing methods. Albedo, normal, and irradiance evaluations are conducted on the Hypersim test set. Roughness and metallicity are evaluated on the Evermotion test set.

Method	Albedo		Method	Normal	
	PSNR↑	LPIPS↓		PSNR↑	LPIPS↓
Our RGB→X	**17.4**	**0.18**	**Our RGB→X**	**19.8**	**0.18**
Zhu et al. [2022b]	11.7	0.54	Zhu et al. [2022b]	16.5	0.45
Careaga and Aksoy [2023]	13.5	0.34	PVT-normal	18.8	0.30
Kocsis et al. [2023]	12.1	0.41			

Method	Roughness		Metallicity		Irradiance	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
Our RGB→X	**11.2**	**0.52**	**12.1**	**0.44**	**14.1**	**0.22**
Zhu et al. [2022b]	4.4	0.77	2.22	0.82	N/A	N/A
Kocsis et al. [2023]	10.3	0.57	8.63	0.75	N/A	N/A

6.3 Applications

Material replacement. In the top left example of Figure 7, we edit the normal and albedo of the sofa (estimated by RGB→X), and re-synthesize the image with our inpainting X→RGB model, resulting in a fuzzier, bumpier red couch. On the top right, we apply intrinsic estimation to the classic Cornell box image and edit the right wall



(a) Albedo estimation on an image from the IIW real-photo dataset [Bell et al. 2014]. Our RGB→X model clearly outperforms that of Zhu et al. [2022b], while those of Kocsis et al. [2023] and Careaga and Aksoy [2023] provide reasonable estimates. Nevertheless, our results show more plausible white balance and flatness, e.g., correctly predicting that all bed-linen pixels should have identical white albedo.

(b) Albedo and irradiance estimation on an image from the MIT Indoor Scene Recognition dataset. Our RGB→X result is the cleanest and has the flattest albedo regions. The result of Careaga and Aksoy [2023] however better preserves the imperfections in the floor albedo.

(c) Our RGB→X roughness estimation on the IIW dataset outperforms previous methods.

(d) Metallicity estimates on the same image as (c).

(e) Normal estimation on an image from the MIT Indoor Scene Recognition real-photo dataset [Quattoni and Torralba 2009]. Our estimate is similar to that of PVT-normal, with slightly better results on flat areas. Note that our model is specialized to indoor scene training data, unlike PVT-normal.

Figure 4: Real-data comparison of our RGB→X model to previous methods.

albedo to blue. We observe that the color bleeding in the rightmost box is correctly updated. The inpainting mask here includes a larger region, allowing for the color-bleeding correction. In the bottom example, we change the normal and albedo of the original room to edit the floor appearance to a wood floor.

Object insertion. In Figure 1(c), we use our framework to insert new synthetic objects into an RGB image. We render the intrinsic channels of the new objects and composite them into the estimated channels. We use our inpainting X→RGB model with rectangular masks to produce a composite with correct lighting and shadows,

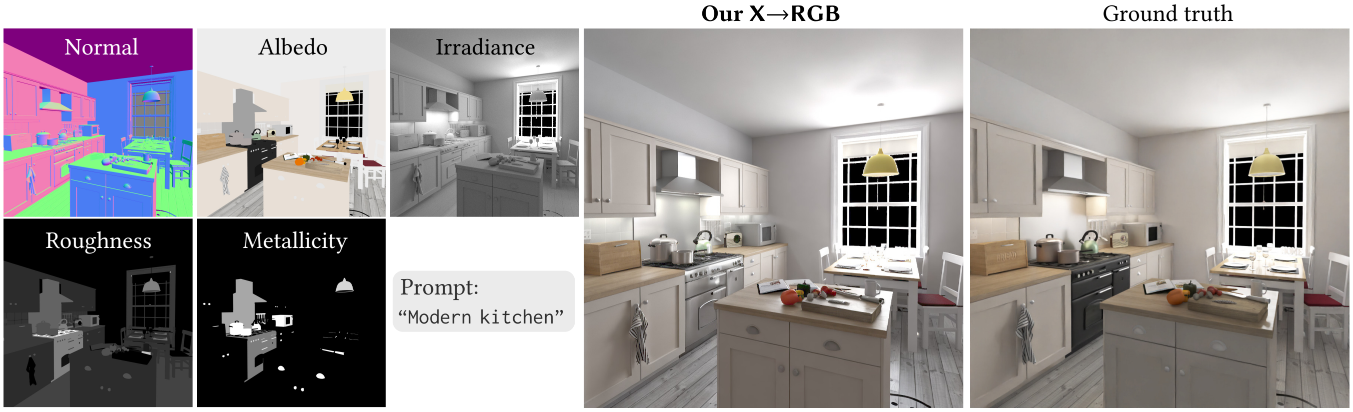
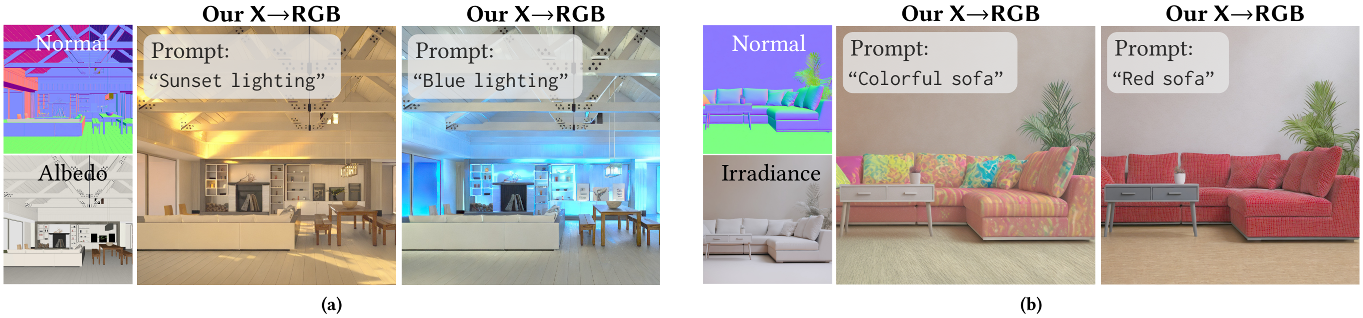


Figure 5: Our X→RGB result on the synthetic kitchen scene [Jay-Artist 2012] which is not part of our training data. We rendered all intrinsic channels, shown on the left, and fed them into the model, along with a text prompt. The result matches the path-traced reference well. There are some differences, e.g., X→RGB makes the stove brighter than the requested albedo, likely because dark metallic materials are rare in the training data.



(a) (b)

Figure 6: X→RGB synthesis given normal and albedo channels only, demonstrating lighting and color-control use of text prompts. (a) Starting from normal and albedo only, we show that the lighting can be controlled by text prompts to some extent. (b) Starting from normal and albedo only, we similarly show the color of objects can be controlled by text prompts to some extent.

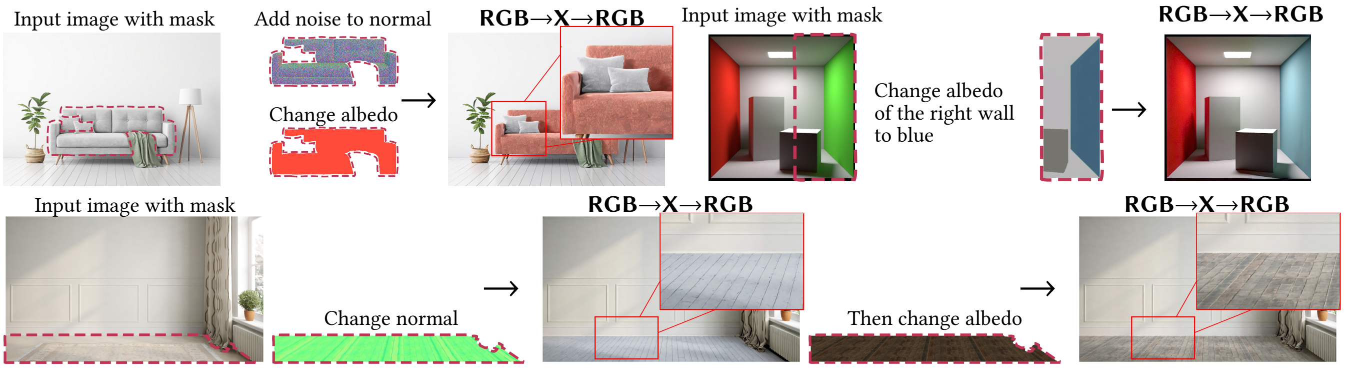


Figure 7: RGB→X and X→RGB models used in combination for material replacement. We show three edit examples. Top left: We change both normal and albedo, resulting in a fuzzier, bumpier red couch. Top right: We edit the right wall albedo of the Cornell box to blue and show that the colour bleeding in the rightmost box is correctly updated. Bottom: We first change the normal to introduce wood planks geometry instead of the original carpet, and then also add wood albedo to edit the floor appearance.

which we finally blend with the original image using a tighter mask. The statue and coffee cart integrate well into the scene.

7 CONCLUSION

In this paper, we explored a unified diffusion framework for intrinsic channel estimation from images (termed RGB→X) and synthesizing realistic images from such channels (X→RGB). Our intrinsic information X contains albedo, normals, roughness, metallicity, and lighting (irradiance). Our RGB→X model matches or exceeds the quality of previous methods, which are specialized to subsets of our intrinsic channels. Our X→RGB model is capable of synthesizing realistic final images, even if we specify only certain appearance properties that should be followed, and give freedom to the model to generate the rest. We show combining both models enables applications such as material editing and object insertion. We believe our work is the first step towards unified diffusion frameworks capable of both image decomposition and rendering, which can bring benefits for a wide range of downstream editing tasks.

Acknowledgements. Ling-Qi Yan is supported by funds from Adobe, Intel, Linctex, Meta and XVerse.

REFERENCES

Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. 1978. Recovering intrinsic scene characteristics. *Comput. vis. syst* 2, 3-26 (1978), 2.

Sean Bell, Kavita Bala, and Noah Snavely. 2014. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 1–12.

Anand Bhattad, Daniel McKee, Derek Hoiem, and David Forsyth. 2024. Stylegan knows normal, depth, albedo, and more. *Advances in Neural Information Processing Systems* 36 (2024).

Reiner Birkl, Diana Wofk, and Matthias Müller. 2023. MiDaS v3.1 – A Model Zoo for Robust Monocular Relative Depth Estimation. *arXiv preprint arXiv:2307.14460* (2023).

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*.

Chris Careaga and Yağız Aksoy. 2023. Intrinsic Image Decomposition via Ordinal Shading. *ACM Trans. Graph.* (2023).

Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. 2023. Generative Models: What do they know? Do they know things? Let's find out! *arXiv preprint arXiv:2311.17137* (2023).

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, 3 (2022), 836–868.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

David Griffiths, Tobias Ritschel, and Julien Philip. 2022. OutCast: Single Image Relighting with Cast Shadows. *Computer Graphics Forum* 43 (2022).

Roger Grosse, Micah K. Johnson, Edward H. Adelson, and William T. Freeman. 2009. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In *International Conference on Computer Vision*. 2335–2342. https://doi.org/10.1109/ICCV.2009.5459428

Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. 2021. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE transactions on knowledge and data engineering* 35, 4 (2021), 3313–3332.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778* (2023).

Jay-Artist. 2012. *Country-Kitchen Cycles.* https://blendswap.com/blend/5156

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. 2023. Intrinsic Image Diffusion for Single-view Material Estimation. In *arxiv*.

Edwin H Land and John J McCann. 1971. Lightness and retinex theory. *Josa* 61, 1 (1971), 1–11.

Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. 2023. Exploiting Diffusion Prior for Generalizable Pixel-Level Semantic Prediction. *arXiv preprint arXiv:2311.18832* (2023).

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs.CV]

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2475–2484.

Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. 2022. Physically-Based Editing of Indoor Scene Lighting from a Single Image. In *ECCV 2022*. 555–572.

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, Miloš Hašan, Ravi Ramamoorthi, and Manmohan Chandraker. 2021. OpenRooms: An Open Framework for Photorealistic Indoor Scene Datasets. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7186–7195. https://doi.org/10.1109/CVPR46437.2021.00711

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. 2024. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5404–5411.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

O. Nalbach, E. Arabadzhiyska, D. Mehta, H.-P. Seidel, and T. Ritschel. 2017. Deep Shading: Convolutional Neural Networks for Screen Space Shading. *Comput. Graph. Forum* 36, 4 (jul 2017), 65–78.

NVIDIA. 2020. *NVIDIA OptiX™ AI-Accelerated Denoiser.* https://developer.nvidia.com/optix-denoiser

Zhaoqing Pan, Weijie Yu, Xiaokai Yi, Asifullah Khan, Feng Yuan, and Yuhui Zheng. 2019. Recent progress on generative adversarial networks (GANs): A survey. *IEEE access* 7 (2019), 36322–36333.

Rohit Pandey, Sergio Orts-Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul E Debevec, and Sean Ryan Fanello. 2021. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.* 40, 4 (2021), 43–1.

Matt Pharr and Greg Humphreys. 2004. *Physically Based Rendering: From Theory to Implementation.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Ariadna Quattoni and Antonio Torralba. 2009. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 413–420.

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 *International conference on machine learning*. PMLR, 8748–8763.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs.CV]

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2022. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 3 (2022).

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 *International Conference on Computer Vision (ICCV) 2021*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. 2022. Nerf for outdoor scene relighting. In *European Conference on Computer Vision*. Springer, 615–631.

Tim Salimans and Jonathan Ho. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512* (2022).

Prafull Sharma, Varun Jampani, Yuanzhen Li, Xuhui Jia, Dmitry Lagun, Fredo Durand, William T. Freeman, and Mark Matthews. 2023. Alchemist: Parametric Control of Material Properties with Diffusion Models. arXiv:2312.02970 [cs.CV]

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Manu Mathew Thomas and Angus G. Forbes. 2018. Deep Illumination: Approximating Dynamic Global Illumination with Generative Adversarial Network. arXiv:1710.09834 [cs.GR]

Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. 2007. Microfacet models for refraction through rough surfaces *(EGSR'07)*. 195–206.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2022. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media* 8, 3 (2022), 415–424.

Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. 2023. Neural Fields meet Explicit Geometric Representations for Inverse Rendering of Urban Scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William AP Smith. 2020. Self-supervised outdoor scene relighting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer, 84–101.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543 [cs.CV]

Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. 2022b. Learning-Based Inverse Rendering of Complex Indoor Scenes with Differentiable Monte Carlo Raytracing. In *SIGGRAPH Asia 2022 Conference Papers*. ACM, Article 6, 8 pages. https://doi.org/10.1145/3550469.3555407

Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. 2022a. IRISformer: Dense Vision Transformers for Single-Image Inverse Rendering in Indoor Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2822–2831.

A NOTE ABOUT TRADITIONAL RENDERING FOR X→RGB

Note that our X→RGB problem cannot be easily solved using traditional rendering. Intrinsic channels (even if all are present) do not contain sufficient information to render a realistic image using traditional techniques, which require full 3D geometry (not just normals) and explicit light/material definitions, including for parts of the scene that are not directly seen by the camera. Screen-space ray tracing / occlusion methods yield only rough approximations, nowhere near the capabilities of our X→RGB model.

Furthermore, when the given intrinsic channels are imperfect or partial, traditional rendering is completely out of the question, but our generative model can still produce reasonable results, possibly controlled with appropriate text prompts.

B IMPLEMENTATION

Training details. We finetune pre-trained Stable Diffusion 2.1 for both RGB→X and X→RGB models. Both models are trained on the InteriorVerse, Hypersim, and Evermotion datasets. We train the X→RGB model additionally on the ImageDecomp dataset, which is constructed from RGB images using our RGB→X model. Both models are trained with a batch size of 256 and using the AdamW optimizer [Loshchilov and Hutter 2017] with a learning rate of 1e−5. We use a random crop of 512 × 512 for training and avoid using a random horizontal flip since it disrupts the camera-space normals. The training of each model takes around 100 hours and the fine-tuning to enable inpainting takes around 20 hours on 8 A100 GPUs.

Inference details. We use the DDIM sampler [Song et al. 2020] with 50 steps for all of our results. We follow the suggestions proposed by Lin et al. [2024] to avoid over-exposure of the generated images. Despite training at 512 × 512 crops, we can run test images at larger resolutions (e.g. 1080p). For comparison, the method of Zhu et al. [2022b] runs on a resolution of 320 × 240 (the default resolution noted in their code).

Classifier-free guidance. Classifier-free guidance (CFG) is commonly used in diffusion models to improve text prompt alignment. For X→RGB, we use classifier-free guidance (CFG) similar to InstructPix2Pix [Brooks et al. 2023]; for RGB→X, we do not use CFG, since we found that it impairs the quality of the RGB→X model and does not provide any benefits, since we do not use text prompts for this model in the usual sense.

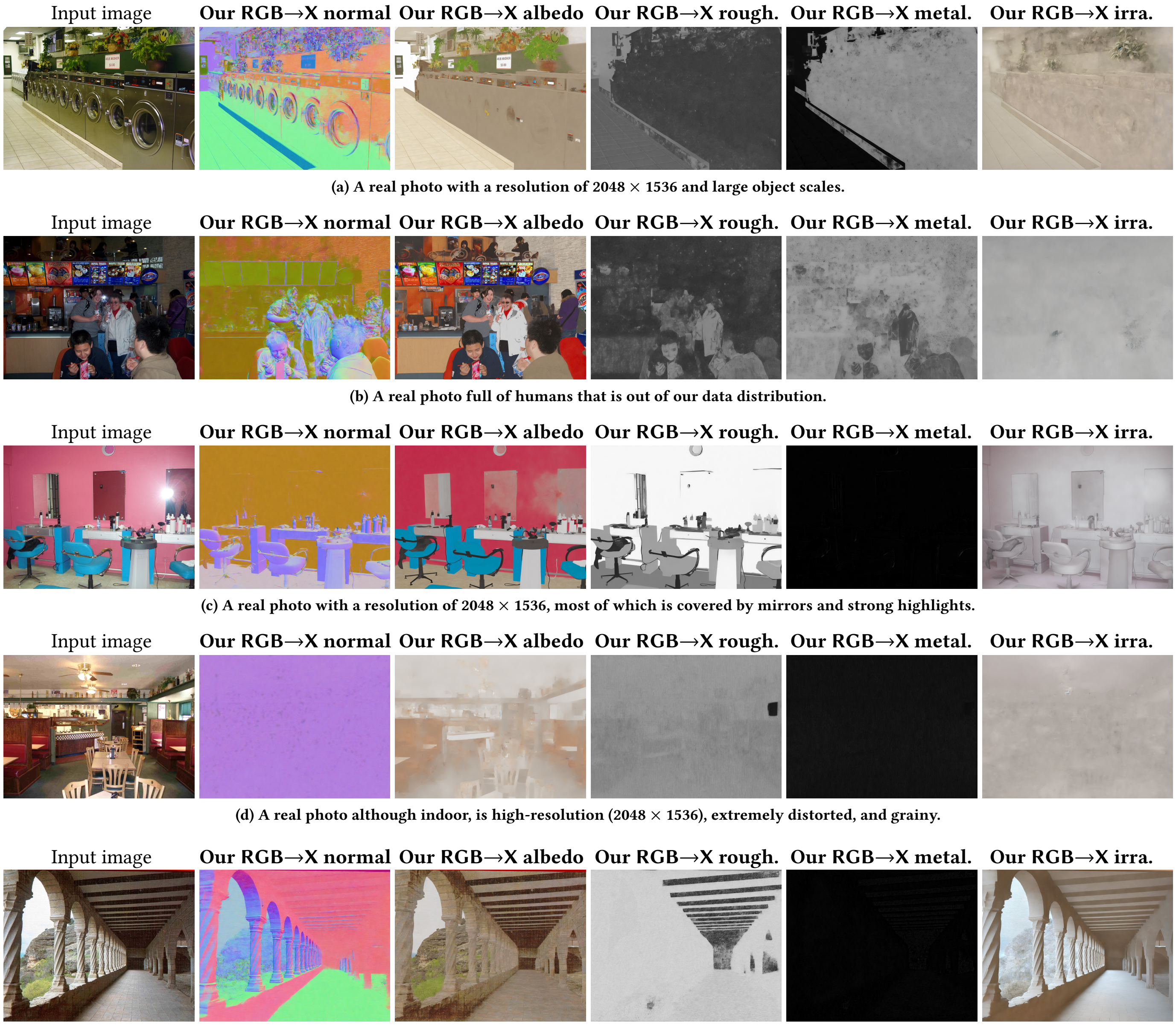
C DISCUSSION AND LIMITATIONS

As our model relies on synthetic dataset for training, different challenges arise. In particular we find the various datasets to present their respective flaws. For example Hypersim sometimes bakes shading into its albedo, and InteriorVerse provides metallic and roughness parameters that are not reliable. Further, the available renderings tend to be noisy and the data often presents aliasing artifacts. While we try to take this into account with our heterogeneous training approach, higher quality, consistent datasets would be beneficial for improved quality. As scene materials datasets tend to be focused on interior scenes, they encode significant bias such as green color being most often for plants, or the fact that e.g. wooden curtains are not common, potentially limiting the editing freedom. Finally, the dropout rate and the probability of picking data from different sources during training can affect the resulting models, guiding them towards preferring some kinds of inputs over others.

In most of these cases, larger, more diverse data would be beneficial in reducing these issues.

In recent generative models a trade-off exists between diversity and adherence to input condition, which can be controlled with Classifier-Free Guidance (CFG). CFG however does not work in our context, and it would be interesting to define such a control mechanism.

As our networks are trained on 512 × 512 resolutions, we can process larger images, but find that quality degrades beyond a 2K resolution; perhaps a coarse-to-fine approach could be used to handle even larger images.



(a) A real photo with a resolution of 2048 × 1536 and large object scales.

(b) A real photo full of humans that is out of our data distribution.

(c) A real photo with a resolution of 2048 × 1536, most of which is covered by mirrors and strong highlights.

(d) A real photo although indoor, is high-resolution (2048 × 1536), extremely distorted, and grainy.

(e) An outdoor real photo, which is out of the our data distribution; this one also contains weird blocking artifacts.

Figure 8: Failure cases of our X→RGB model on the MIT Indoor Scene Recognition real photo dataset [Quattoni and Torralba 2009].

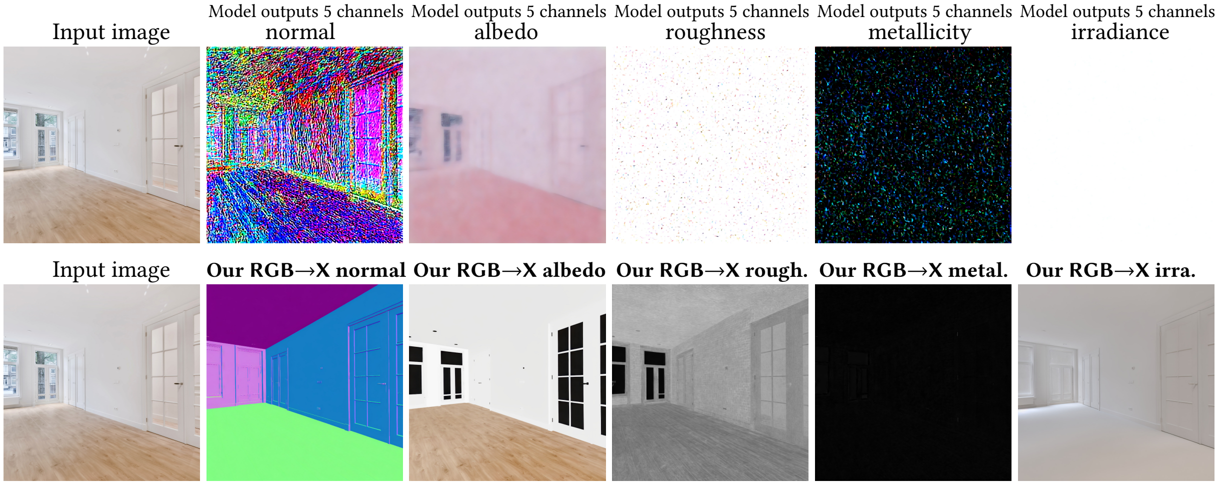


Figure 9: The model in the top row outputs a larger latent vector for 5 channels (normal, albedo, roughness, metallicity, and irradiance) at once. However, we find this model is hard to train and performs poorly even after 100 epochs of training.

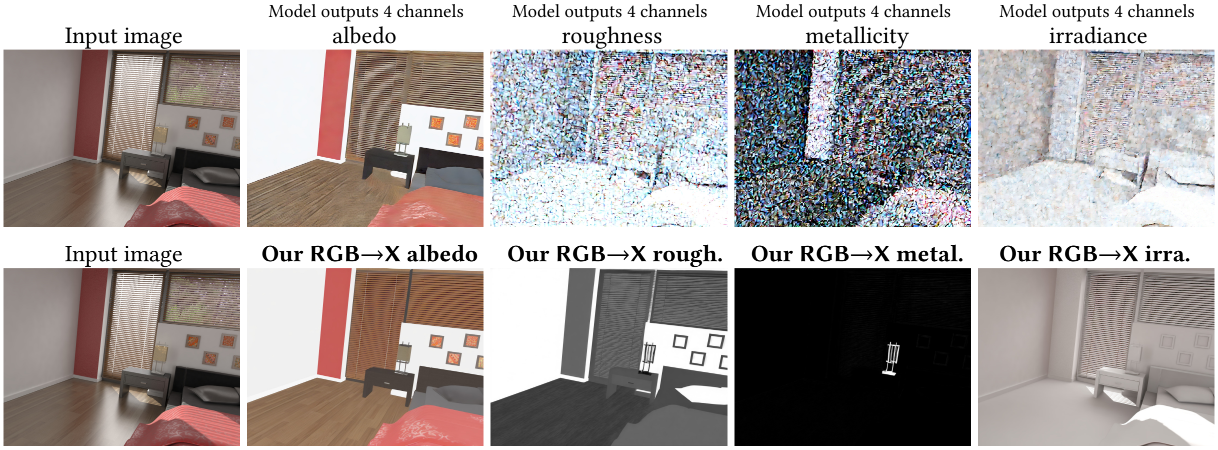


Figure 10: The model in the top row outputs a larger latent vector for 4 channels (albedo, roughness, metallicity, and irradiance) at once; compared to the model outputs 5 channels in Figure 9 (top), this model starts to generate reasonable results in albedo after 100 epochs of training. Still, it performs poorly on the other channels.

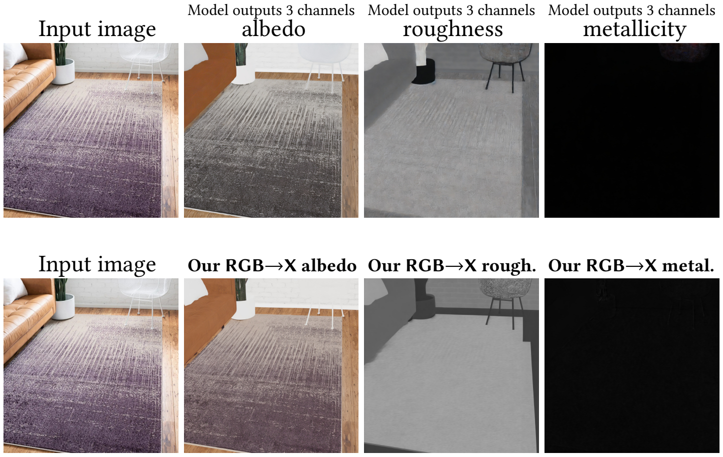


Figure 11: The model in the top row outputs a larger latent vector for 3 channels (albedo, roughness, and metallicity). Compared to models in Figure 9 and Figure 10, this model can produce flat and clean albedo channel and decent roughness and metallicity channel. However, compared to our model trained with the same number of epochs, this model produce a color shift in albedo channel, and distortions in roughness channel.