Generative Models: What do they know? Do they know things? Let's find out!

Xiaodan Du[1] Nicholas Kolkin[2] Greg Shakhnarovich[1] Anand Bhattad[1]
[1]Toyota Technological Institute at Chicago [2]Adobe
https://intrinsic-lora.github.io/

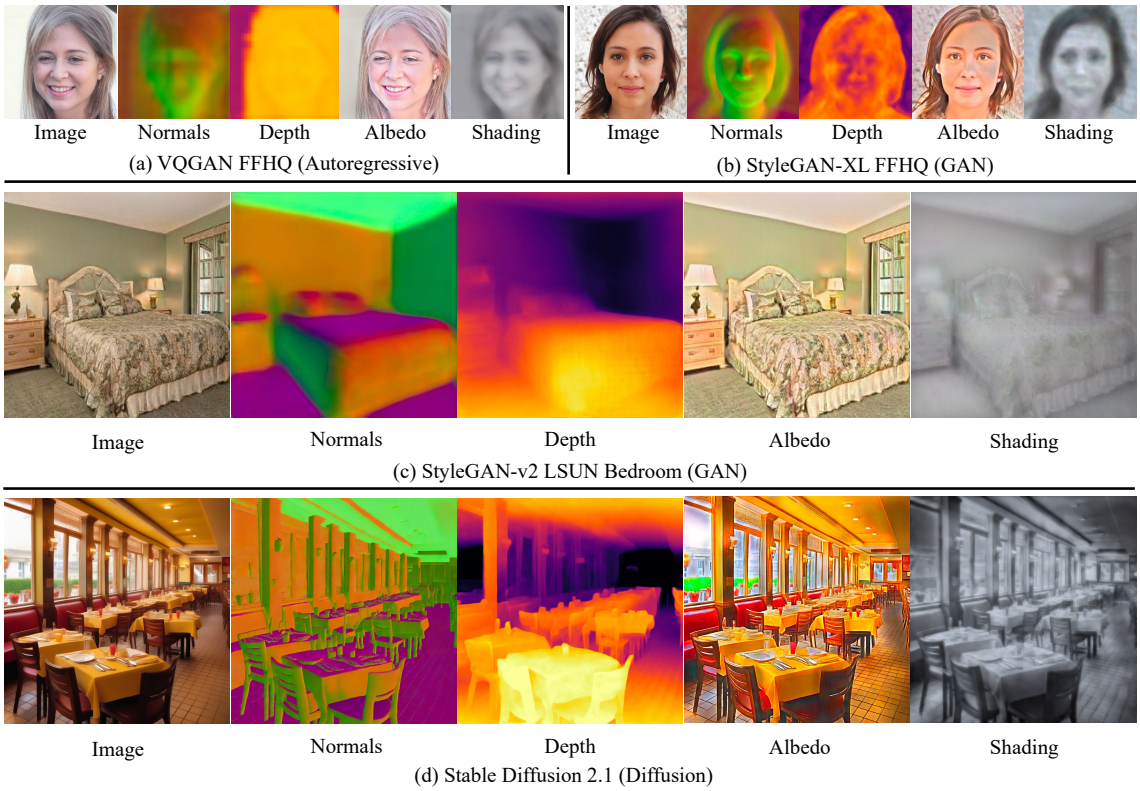


Figure 1. We introduce INTRINSIC LORA (I-LORA) that reveals the hidden capabilities of various generative models. This includes VQGAN (a), StyleGAN-XL (b), StyleGAN-v2 (c), and Stable Diffusion (d). Our approach uses a low-rank adaptation (LoRA) method to modulate key feature maps. These include attention layers in VQGAN and Stable Diffusion and affine layers in StyleGAN. This modulation allows the models to reveal intrinsic properties such as normals, depth, albedo, and shading. I-LORA effectively uses the existing decoder for intrinsic image extraction, which was previously utilized for RGB image generation. It does so without the need for new layers, showcasing the deep, inherent understanding these models have of complex scene intrinsics.

Abstract

Generative models have been shown to be capable of synthesizing highly detailed and realistic images. It is natural to suspect that they implicitly learn to model some image intrinsics such as surface normals, depth, or shadows. In this paper, we present compelling evidence that generative models indeed internally produce high-quality scene intrinsic maps. We introduce INTRINSIC LORA *(*I-LORA*), a universal, plug-and-play approach that transforms any generative model into a scene intrinsic predictor, capable of extracting intrinsic scene maps directly from the original generator network without needing additional decoders or fully fine-tuning the original network. Our method employs a Low-Rank Adaptation (LoRA) of key feature maps, with newly learned parameters that make up less than 0.6% of the total parameters in the generative model. Optimized with a small set of labeled images, our model-agnostic approach adapts to various generative architectures, including Diffusion models, GANs, and Autoregressive models. We show that the scene intrinsic maps produced by our method compare well with, and in some cases surpass those generated by leading supervised techniques.*

1. Introduction

Generative models are capable of producing high-quality images that can be mistaken for real-world photographs. These models display a seemingly profound understanding of the world, capturing the nuances of realistic object placement, appearance, and lighting conditions. However, the mechanisms by which these models acquire such detailed knowledge remain largely unexplored. One of the most pressing questions in this context is: What kind of knowledge do generative models rely on to produce such high-quality images? Are they manipulating abstract, high-level representations of the world, working with more physical, concrete scene representations, or perhaps a combination of both?

Recent work has begun to shed light on this question. For instance, Bhattad et al.[8] demonstrated that StyleGAN can encode important scene intrinsics like depth and normals. Similarly, Zhan et al.[60] showed that diffusion models can understand 3D scenes in terms of geometry and shadows. Chen et al.[11] found that Stable Diffusion's internal activations encode 3D depth and saliency maps that can be extracted with linear probes. Three independent groups [22, 37, 51] found correspondences in generative diffusion models without explicit supervision. However, these studies are often model-specific and do not address whether these capabilities are inherent to all large-scale generative models or are a result of specific architectural choices.

In this paper, our goal is to understand the underlying knowledge present in all types of generative models. We employ Low-Rank Adaptation [25] (LoRA) as a unified approach to extract scene intrinsic maps—namely, normals, depth, albedo, and shading—from different types of generative models. Our method, which we have named as INTRINSIC LORA (I-LORA), is general and applicable to diffusion-based models, StyleGAN-based models, and autoregressive generative models. Importantly, the additional weight parameters introduced by LoRA constitute less than 0.6% of the total weights of the pretrained generative model, serving as a form of feature modulation that enables easier extraction of latent scene intrinsics. By altering these minimal parameters and using as few as 250 labeled images, we successfully extract these scene intrinsics.

Why is this an important question? Our motivation is three-fold. First, it is scientifically interesting to understand whether the increasingly realistic generations of large-scale text-to-image models are correlated with a better understanding of the physical world, emerging purely from applying a generative objective on a large scale. Second, rooted in the saying "vision is inverse graphics" – if these models capture scene intrinsics when generating images, we may want to leverage them for (real) image understanding. Finally, analysis of what current models do or do not capture may lead to further improvements in their quality.

How is what we do related to fine-tuning or linear probing? Previous approaches that explore extracting intrinsics from trained generative models include fine-tuning the models [63] or (usually linear) probing [11]. Fine-tuning a model, using a dataset of images paired with target scene intrinsic maps, yields a new model, no longer capable of generating images. Since all or most of the parameters have changed, it is unclear whether the fine-tuned model's ability to produce, say, scene depth is also inherent in the original model. In contrast, our I-LORA approach only introduces a tiny number of extra parameters, effecting minimal change and leaving the original model easily accessible.

Model	Pretrain Type	Domain	Normal	Depth	Albedo	Shading
VQGAN [16]	Autoregr.	FFHQ	∼	∼	✓	✓
SG-v2 [31]	GAN	FFHQ	✓	∼	✓	✓
SG-v2 [59]	GAN	LSUN Bed	✓	✓	✓	✓
SG-XL [50]	GAN	FFHQ	✓	∼	✓	✓
SG-XL [50]	GAN	ImageNet	×	×	×	×
SD-UNet [46]	Diffusion	Open	✓	✓	✓	✓
SD [46]	Diffusion	Open	✓	✓	✓	✓

Table 1. Summary of scene intrinsic extraction capabilities across different generative models without changing generator head. ✓: Intrinsics can be extracted with high quality. ∼: Intrinsics cannot be extracted with high quality. ×: Intrinsics cannot be extracted.

We do not construct additional "layers" on top of model activations, like the probing methods. Instead, we learn to leverage the existing model along with the "adaptor" (LoRA) so that it produces a scene intrinsic map. In this we take advantage of the fact that for all the intrinsics we study, these maps can be represented as an image with up to three channels– something the generative models are already set up to produce, albeit the nature of these new images is different, requiring the LoRA feature modulation.

What specifically do we study? We conduct experiments with a variety of generative models: diffusion models [24, 32, 46], GANs [29, 31] and autoregressive models [16]. While diffusion models such as StableDiffusion [46] or Imagen [48] have perhaps received most limelight recently, the most recent generation of other types of models such as GigaGAN [27], CM3leon [58], and Parti [57] appear able to produce images of similar quality. We believe it is important to study all of these, and even more important to consider a general framework that would be applicable to all these models – and perhaps to others yet to be proposed.

We evaluate the quality of the intrinsics extracted from the generative models on both real and synthetic images. As training targets, we use predictions by state of the art specialized networks trained on large datasets as pseudo ground truth. When available, we also compare the predicted intrinsics to real ground truth. Finally, we inspect the correlation between the generator's image quality and the quality of the intrinsics that can be extracted from it.

Our findings reveal that *all* types of the generative models we study contain rich information about scene intrinsics that can be easily extracted using LoRA. A summary of our findings is in Table 1, with more details in Section 4.

Figure 2. Scene intrinsics from different generators – VQGAN, StyleGAN-v2, and StyleGAN-XL – trained on FFHQ dataset: The first column shows the original synthetic images produced by each model. The subsequent columns show the scene intrinsics corresponding to surface normals, depth, albedo, and shading, as extracted by a SOTA-non-generative model and our methodology (I-LoRA). For surface normals, the images highlight the models' ability to infer surface orientations and contours. The depth maps display the perceived distances within the images, with warmer colors indicating closer objects and cooler colors representing further ones. Albedo maps isolate the intrinsic colors of the subjects, removing the influence of lighting and shadow. Finally, the shading maps capture the interplay of light and surface, showing how light affects the appearance of different facial features. Intriguingly, there appears to be a correlation between the fidelity of the generated images and the accuracy of the surface normals extracted. This suggests that the more sophisticated the image generation capabilities of a model, the more precisely it can replicate the nuances of real-world physics in its scene intrinsics.

Our work takes a significant step towards demystifying the intrinsic capabilities of generative image models. It also may open new avenues for leveraging generative models in various applications, from computer vision to graphics and beyond, where these intrinsics are important. In summary, our main contributions are:

- Showing that the same INTRINSIC LoRA method is sufficient for general scene intrinsic extraction across various generative models.
- Demonstrating high-quality scene intrinsic extraction outperforming SOTA supervised dense prediction models on some tasks, with minimal addition of parameters (less than 0.6% of the number of the original model parameters in all cases) and as little as 250 labeled images.
- Demonstrated correlation between the quality of scene intrinsic maps extracted and the image generation quality of a model. This provides an alternative perspective for evaluating generative models.

2. Related Work

Generative Models: Generative Adversarial Networks (GANs) [20] have been widely used for generating realistic images. Variants like StyleGAN [29], StyleGAN2 [31] and GigaGAN [27] have pushed the boundaries in terms of image quality and control over the generated content. Some works have explored the interpretability of GANs [4, 8], but few have delved into their ability to capture scene intrinsics.

Diffusion models, such as Denoising Score Matching [55] and Noise-Contrastive Estimation [21], have been used for generative tasks and are perhaps the most popular at the

moment [24, 32, 46]. These models have been shown to understand complex scene intrinsics like geometry and shadows [60], but their generalizability across different scene intrinsics is less explored.

Autoregressive models like PixelRNN [53] and PixelCNN [52] generate images pixel-by-pixel, offering fine-grained control but at the cost of computational efficiency. More recently, VQ-VAE-2 [45] and VQGAN [16] have combined autoregressive models with vector quantization to achieve high-quality image synthesis. While these models are powerful, their ability to capture and represent scene intrinsics has not been thoroughly investigated.

Scene Intrinsics Extraction: Barrow and Tenenbaum [3] highlighted several important fundamental scene intrinsics that include depth, albedo, shading, and surface normals. Several works have focused on extracting scene intrinsics like depth and normals from images [5, 14, 15, 28, 35, 44] using labeled annotated data. Labeled annotations of albedo and shading are hard to find and as the recent review in [17] shows, methods involving little or no learning have remained competitive until relatively recently. However, these works often rely on supervised learning and do not explore the capabilities of generative models in this context.

Many recent studies have utilized generative models such as [1, 2, 26, 33, 40, 49, 56, 62, 63] as pretrained feature extractors or scene prior learners. These models use generated images to enhance downstream discriminative models, fine-tune the original generative model for a new task, learn new layers, or develop new decoders to produce desired scene intrinsics. InstructCV [19] executes different computer vision tasks via natural language instructions, abstracting task-specific design choices. However, it requires retraining of the entire diffusion model. It's still unclear whether the original models inherently capture important scene intrinsics information as implicit knowledge or not.

Knowledge in Generative Models: Several studies have explored the extent of StyleGAN's knowledge, particularly in the context of 3D information about faces [42, 61] provide substantial evidence of StyleGAN's capability in this area. Further research has demonstrated that manipulating offsets in StyleGAN can lead to effective relighting of images [6], as well as the extraction of scene intrinsics [8].

Chen et al.[11] found that the internal activations of the LDM encode linear representations of both 3D depth data and a salient-object / background distinction. Recently, [22, 37, 51] found correspondence emerges in image diffusion models without any explicit supervision.

Self-supervised models like DINO [10, 12, 41] focus on learning useful representations without using labeled data. Although not generative, these models serve as a relevant baseline for understanding the quality of scene intrinsics that can be extracted from learned representations.

LoRA (Low-Rank Adaptation) is a technique originally used to reduce the cost of fine-tuning large language models for downstream tasks [25]. The approach involves freezing the pre-trained model weights and introducing trainable low-rank decomposed matrices into specific layers of model architecture. These matrices are the only components updated during task-specific optimization. This results in a significant reduction in the number of trainable parameters. LoRA has been used for various personalization applications of image generators [47]. In contrast, we use LoRA as a unified and efficient method for extracting scene intrinsics across various types of generative models. It does this by identifying key feature maps in each of them to modulate – attention layer in diffusion models, linear affine layers in StyleGAN and convolutional attention layer in VQGAN.

3. Methods

In the most general formulation, a generative model G maps noise/conditioning information z to an RGB image $G(z) \in \mathbb{R}^{H\times W\times 3}$. We seek to augment G with a small set of parameters θ that allow us to produce, using the same architecture as G, an image-like object with up to three channels, representing scene intrinsics like surface normals.

The learning framework: We learn to extract intrinsics in a supervised fashion. Since in most cases (some domains, generated images) we do not have ground truth intinsics, we use state of the art models (trained on large datasets) to predict "pseudo ground truth" maps, e.g., estimated depth for an image, and treat these as a target for the predictions of G_θ.

In order to optimize θ of G_θ we employ a pseudo-ground truth predictor Φ (e.g., a network trained to predict depth map from an image), leading to the objective function:

$$\min_\theta \mathbb{E}_z[d(G_\theta(z), \Phi(G(z)))], \tag{1}$$

where d is a distance metric that depends on the intrisic we wish to learn.

Diffusion models require a more tailored treatment, since they are effectively image-to-image and not noise-to-image (since during inference they repeatedly get a noisy image as input). Thus instead of conditioning noise z we feed an image x(generated or real) to a diffusion model G. In this case, given a real image x, our objective function becomes $\min_\theta \mathbb{E}_x[d(G_\theta(x), \Phi(x))]$.

We describe the metric d and the pseudo-ground truth predictors Φ used for each intrinsic in Section 4.1.

Low-Rank Adaptation Low-Rank Adaptation (LoRA) is a parameter-efficient technique for adapting pre-trained neural networks to novel tasks by introducing a low-rank weight matrix W^*. Initially devised for language models [25], LoRA has since been extended to image generation models where it is primarily used to enable the generation of new characters, objects, or styles [47].

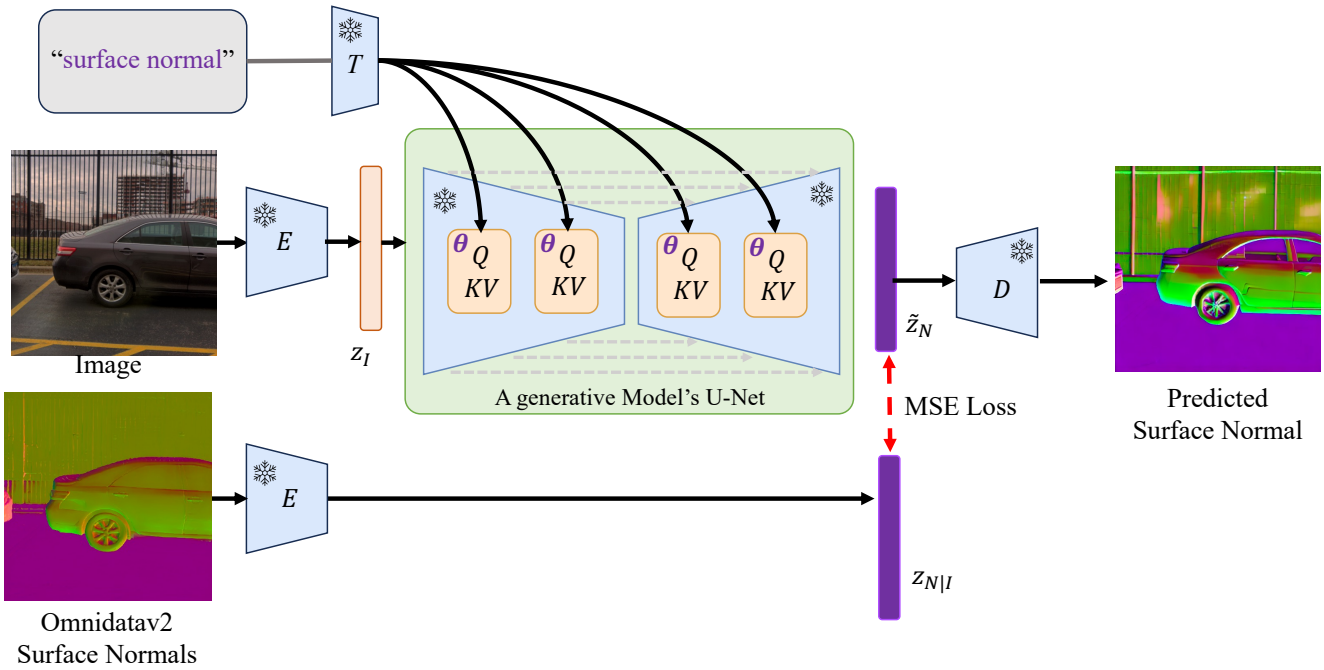


Figure 3. Overview of our INTRINSIC LORA applied to a single-step diffusion model. We fine-tune low-rank matrices corresponding to key feature maps—specifically, attention matrices—to elicit scene intrinsics. Distinct low-rank adaptors, indicated by color-coded θ (e.g., purple for surface normals), are optimized for each intrinsic. We leverage a limited set of labeled examples to empower generative models to directly extract scene intrinsics using the original decoder, circumventing the need for specialized decoders or comprehensive model retraining.

In LoRA, W^* is designed to have a lower rank than the original weight matrix $W \in \mathbb{R}^{d_1 \times d_2}$, achieved by factorizing W^* into two smaller matrices, $W^* = W_u^* W_l^*$, $W_u^* \in \mathbb{R}^{d_1 \times d^*}$ and $W_l^* \in \mathbb{R}^{d^* \times d_2}$, where $d^* \ll \min(d_1, d_2)$.

The output o for an input activation a is then given by:

$$o = Wa + W^*a = Wa + W_u^* W_l^* a. \tag{2}$$

To preserve the original model's behavior at initialization, W_u^* is set to zero. We next describe how we leverage LoRA modules to extract intrinsics from Diffusion models, GANs, and Autoregressive models.

When adapting **diffusion models** for intrinsic extraction we learn I-LORA adaptors on top of the cross-attention and self-attention layers. We treat the UNet as a dense prediction model, applying it once to an RGB input with the goal of getting an intrinsic map as output. We find this gives the best quantitative results. The text input depends on the intrinsic and is simply "surface normal, "depth", "albedo", or "shading". The timestep input is set to T=1.0 (i.e. full noise), we also tested T=0.0 and it yielded equivalent results. An overview of I-LORA is in Figure 3.

When adapting **GANs** we learn I-LORA modules on top of the affine layers projecting from w-space to s-space.

When adapting **VQGAN, an autoregressive model,** we learn I-LORA modules on top of the the convolutional attention layers in the decoder.

4. Experiments

4.1. Implementation Details

Notably the compute, parameter, and data requirements for learning these adaptors is extremely light. In all cases we use I-LORA adaptors with rank 8 which add less than 0.6% to the original parameters. We provide detailed training hyperparameters in the supplement.

To generate pseudo ground truth for depth we use ZoeDepth [5] as the predictor Φ in Equation (1). For surface normals Φ is OmniDataV2-Normal [14, 28]. For Albedo and Shading Φ is Paradigms [7, 17].

For SG2, SGXL and VQGAN, d in Equation1 is

$$d(x, y) = 1 - cos(x, y) + \|x - y\|_1 \tag{3}$$

for normal and MSE for other intrinsics. For latent diffusion based methods, there isn't a clear physical meaning to the relative angle of latent vectors in encoded normal maps, so we use the standard objective of MSE for all intrinsics.

4.2. Synthetic Image Experiments

We begin I-LORA by extracting intrinsics from **generated** images. We consider a variety of generative models (StyleGAN-v2, StyleGAN-XL, VQGAN) trained on a variety of datasets (FFHQ, LSUN Bedrooms, ImageNet). For each model trained on a particular dataset, we train

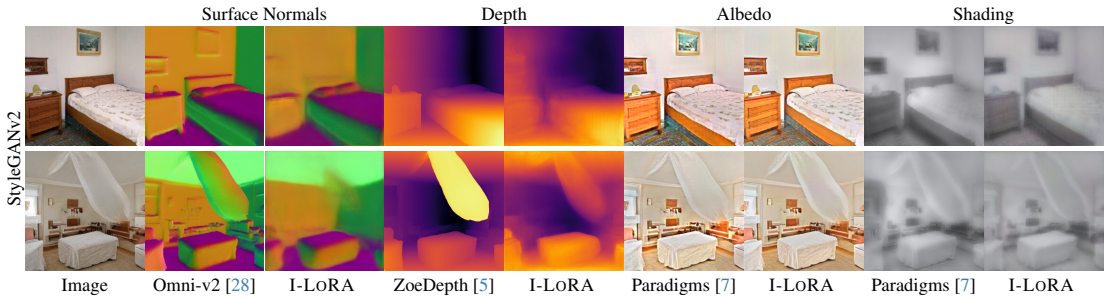


Figure 4. Scene intrinsics extraction from Stylegan-v2 trained on LSUN bedroom images.

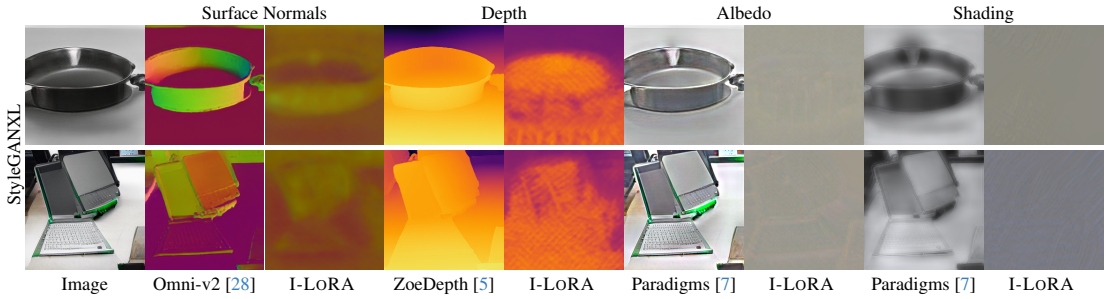


Figure 5. StyleGAN-XL trained on ImageNet. Top: pan, bottom: laptop, with the corresponding scene intrinsics (pseudo ground truth and extracted) alongside. The surface normals and depth maps, while capturing the basic shape and volume, lack precise detail and exhibit artifacts. The albedo maps fail to consistently separate texture from lighting, and the shading maps do not fully capture the nuanced interplay of light and surface. These difficulties are correlated with the overall worse realism of the generated images.

Model	Pretraining Type	Domain	LoRA Param.	Surface Normal			Depth		Albedo	Shading
				Mean Error°↓	Median Error°↓	L1 Error$_{\times 100}$↓	RMS$_{\times 100}$↓	$\delta < 1.25_{\times 100}$↑	RMS$_{\times 100}$↓	RMS$_{\times 100}$↓
VQGAN	Autoregressive	FFHQ	0.18%	19.97	20.97	16.33	1.819	62.33	0.345	0.106
StyleGAN-v2	GAN	FFHQ	0.57%	16.78	19.22	13.72	1.530	90.74	0.283	0.110
StyleGAN-XL	GAN	FFHQ	0.29%	14.87	18.22	12.27	1.337	93.87	0.287	0.125
StyleGAN-v2	GAN	LSUN Bedroom	0.57%	13.24	23.57	10.91	0.897	66.88	0.270	0.074
StyleGAN-XL	GAN	ImageNet	0.29%	24.09	25.52	19.44	2.175	38.38	1.065	0.119
AUGUNET (multi step)	Diffusion	Open	0.18%	21.41	28.57	17.39	2.042	41.21	0.881	0.099
SD UNet (single step)	Diffusion	Open	0.18%	16.63	23.64	13.69	1.179	52.59	0.487	0.118

Table 2. Quantiative analysis of scene intrinsics extraction performance on generated images. We compare with pseudo ground truths from Omnidata-v2 for surface normals, ZoeDepth for depth, Paradigms for albedo and shading. Metrics include mean angular error, median angular error, and L1 error for surface normals; rms and $\delta < 1.25$ for depth; rms for albedo and shading. We also include results of SD UNet and AUGUNET (described in details in Sec 5) on 1000 synthetic images with various prompts for completeness.

I-LoRA adaptors to produce surface normals, depth, albedo, and shading aligned with the RGB images produced by the original generator. We generally find the I-LoRA adaptors can successfully produce predictions that approximately match the pseudo ground-truth quantitatively (Table 2), and qualitatively produce compelling results (Figures 2, 4).

We use the 256^2 resolution checkpoints for VQGAN, StyleGAN-v2 and StyleGAN-XL trained on FFHQ dataset; with FID [23] scores of 9.6 [16], 3.62 [30] and 2.19[1], respectively. Our qualitative results in Table 2 indicate a decreasing FID (i.e. better image generation) correlates with better intrinsic prediction.

One exception to this is StyleGAN-XL trained on ImageNet, which we find to produce reasonable quantitative metrics, but poor qualitative results. We attribute this to the underlying quality of the generative model, noting that many samples from this model could never be mistaken for natural images (Figure 5). This generally supports our finding that the stronger the generative model, the higher the quality of the extracted intrinsics.

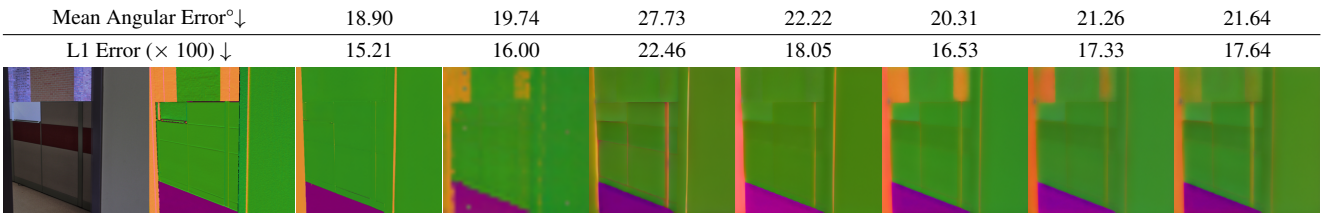
4.3. Real Image Experiments

While I-LoRA results on synthetic images and pseudo-ground truth are suggestive, they are difficult to judge objectively. Fortunately, diffusion models are not only powerful image generators, their structure as image-to-image models makes it straightforward to apply to real images. We adapt Stable Diffusion's (SD's) Unet to extract intrinsics but apply it once as a dense predictor, rather than iteratively as

[1]https://github.com/autonomousvision/stylegan-xl

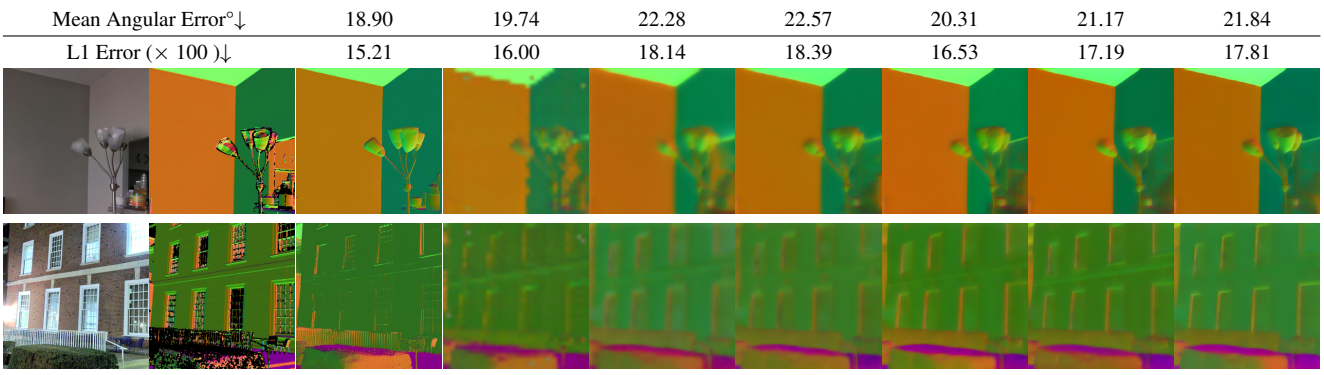
Model	Pretraining Type	Surface Normal			Depth	
		Mean Error°↓	Median Error°↓	L1 Error$_{\times 100}$↓	RMS$_{\times 100}$↓	$\delta < 1.25_{\times 100}$↑
Omnidata-v2 [28]/ZoeDepth [5]	Supervised	18.90	13.36	15.21	2.693	47.56
DINOv2	Non-Generative	19.74	13.72	16.00	2.094	44.32
AUGUNET (multi step)	Diffusion	23.74	19.08	19.31	2.651	43.19
SD UNet (single step)	Diffusion	20.31	12.54	16.53	2.046	44.90

Table 3. Quantitative analysis of scene intrinsic extraction performance across different models on real images. See caption of Tab 2 for details on pseudo-ground truth and metrics.



(a) Real (b) GT (c) Omni-v2 [28] (d) DINOv2[41] (e) 250 (f) 1000 (g) 4000 (h) 16000 (i) 24895

Figure 6. Ablation study on the number of training samples using a single step SD UNet. We use surface normal prediction to illustrate. Column (c) is the Omnidata-v2 pseudo label we use for training. Numbers below columns (e) to (i) are the number of samples used during training. The results demonstrate that even with a limited dataset comprising a few hundred images, the pretrained stable diffusion model can discern substantial surface normal details. Notably, with a dataset size of merely 4000 samples, our approach accurately captures challenging areas, such as the top left distant walls in (a). We started all models with SD v1-5 and LoRA rank=8.



(a) Real (b) GT (c) Omni-v2 [28] (d)DINOv2 [41] (e) Rank = 2 (f) Rank = 4 (g) Rank = 8 (h) Rank = 16 (i) Rank = 32

Figure 7. Ablation study–LoRA's Rank. Numbers under columns (e) to (i) are the numbers of ranks of LoRA. 8 achieves memory efficiency and gives good results. All models are trained with SD v1-5 and 4000 samples.

in diffusion. We train I-LORA adaptors for surface normal and depth extraction using a random subset of 4000 training pairs from the DIODE [54] training set. While we use pseudo ground truth during training, DIODE provides real ground truth we can use for quantitative evaluation.

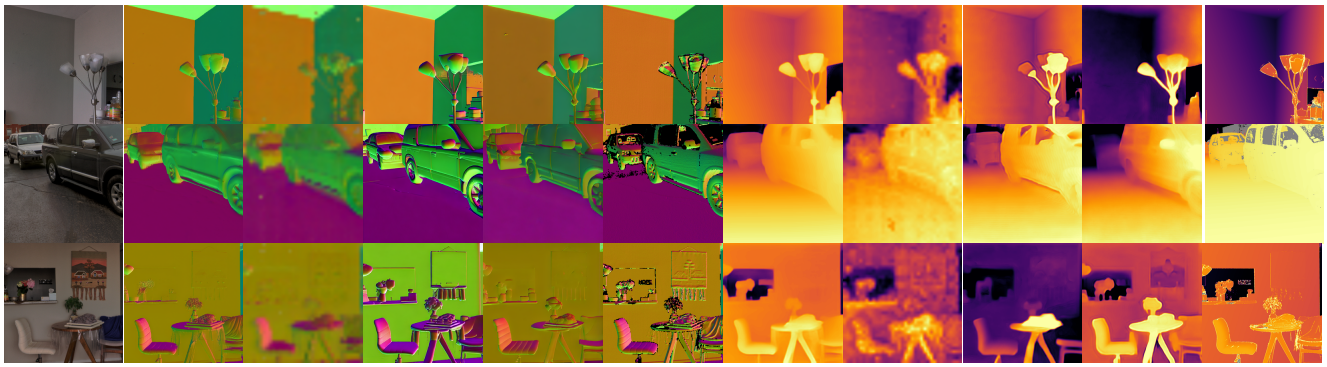
We find that not only I-LORA adaptors capable of roughly matching the performance of Φ (which provided the training signal) while using far less data, parameters, and training time; but they even surpass Φ in several metrics (Table 3). A potential explanation is that the intrinsic implicitly learned by the generative model is actually more accurate than that learned by the supervised SOTA predictor Φ.

4.4. Ablation Studies

In our investigation, we conduct a series of ablation studies focusing on our single step SD UNet model, which has the highest quantitative performance. We explore the impact of the number of labeled examples on the model's ability to extract scene intrinsics (Figure 6), compare multiple versions of stable diffusion models to further explore how improved generation quality affects extrinsic extraction (Figure 10, and determine the optimal rank for our low-rank adaptations (Figure 7). In short these ablation studies reveal that higher quality generative models lead to better intrinsic extraction, and that the quality of intrinsic extraction saturates with fairly few training samples (4000) and low rank LoRAs (rank 8).

4.4.1 Comparison with DINOv2

A natural question is how do the features learned by generative models compare with those from other self-supervised foundation models. To provide a preliminary benchmark we



(a) Real Image (b) Omni-v2 (c) DINOv2 (d) AUGUNET1.5 (e) AUGUNET (f) GT (g) ZoeDepth (h) DINOv2 (i) AUGUNET1.5 (j) AUGUNET (k) GT

Figure 8. Detailed Scene Intrinsic Extraction with Improved Diffusion Techniques from Our AUGUNET models: We show scene intrinsics derived from generative models alongside supervised counterparts. AUGUNET1.5 is the same as AUGUNET except it uses SD1.5 and does not use Zero SNR strategy. AUGUNET1.5 presents sharper details, especially in complex areas – thin structures like lamp stand and car. AUGUNET, on the other hand, illustrates a significant improvement in reducing color shifts while maintaining detail sharpness, as seen in the comparison with ground truth (GT) data in the last row.

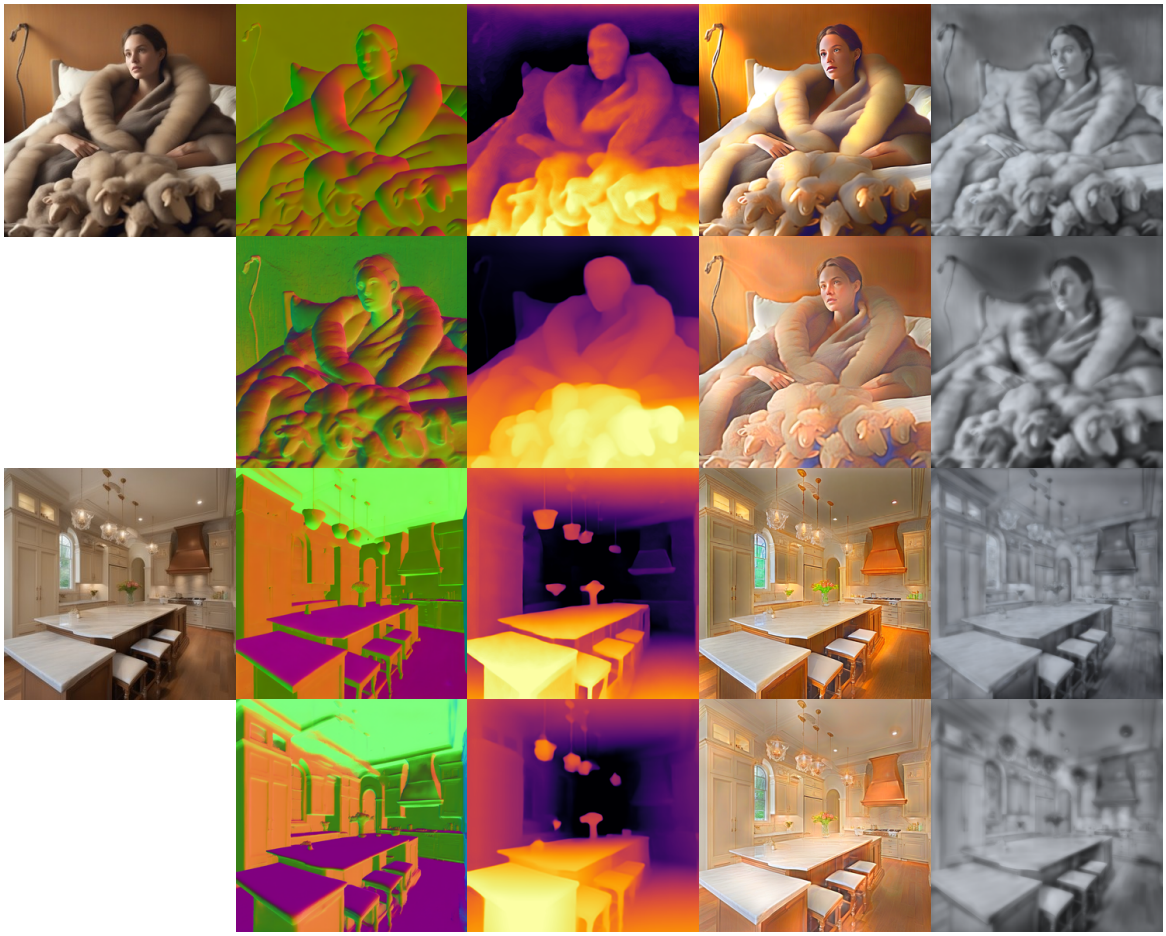


Figure 9. Results of AUGUNET models applied on unseen 1024^2 synthetic images. For each group, from left to right, we have: image, surface normal, depth, albedo and shading, respectively. The first row contains results from the I-LORA . The second row contains the corresponding pseudo ground truth from off-the-shelf SOTA methods.

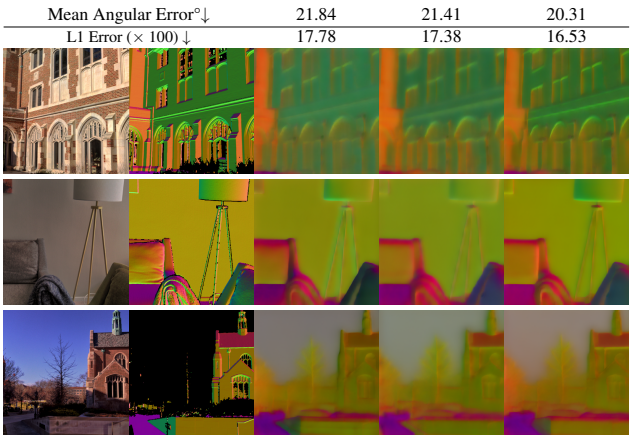


Figure 10. Through an ablation study of various Stable Diffusion versions using a single-step method, we observe a correlation between the quality of the generative model and the accuracy of the extracted scene intrinsics. Our analysis includes Stable Diffusion versions 1.1, 1.2, and 1.5, each demonstrating significant enhancements in image generation capabilities, which in turn improve the fidelity of scene intrinsics extraction.

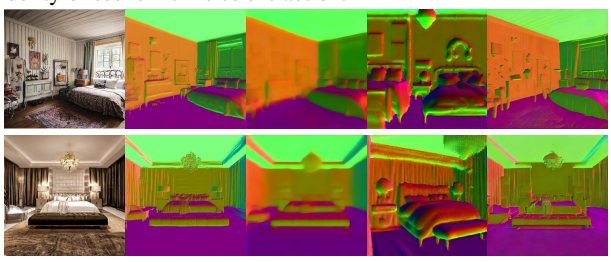


Figure 11. Single-step diffusion (SD UNet) yields satisfactory results, but multiple diffusion steps lead to misalignment in the extracted intrinsics, as shown in the SD1.5 (multi-step) column. The last column, AUGUNET , demonstrates the successful rectification of these misalignments using our image conditioning approach, resulting in well-aligned and coherent scene intrinsics extractions

apply our I-LORA framework to DINOv2 [41] , a SOTA general purpose visual feature extractor. We start with their best and largest model, "giant", and apply rank 8 LoRA modules to all attention layers. In addition, we must learn a small linear head to project DINO output to 3 channels (1 channel in the case of depth). These modules lead to a parameter increase of 0.26%, similar to those added for generative models (Table 2). We train these parameters on DIODE images using the same procedure outlined in Section 4.3, finding that the results are comparable to our best model single-step model based on SD, see Table 3.

5. Towards Improved Intrinsic Extraction

Our findings suggest that Stable Diffusion models inherently capture various scene intrinsics like normals, depth, albedo, and shading, as evidenced by our single-step SD Unet experiments. However, a natural question is: can we enhance the quality of the intrinsics by leveraging the *multi-step* diffusion inference? While multi-step diffusion improves sharpness, in practice we find that it introduces two challenges: 1. intrinsics is misaligned with input, and 2. a shift in the distribution of outputs relative to the ground truth (visually manifesting as a color shift) (see Figure 11).

To address the first challenge, we augment the noise input to the UNet with the input image's latent encoding, as in InstructPix2Pix [9] (IP2P). The second challenge is a known artifact attributed to Stable Diffusion's difficulty generating images that are not with medium brightness [13, 34]. Lin et al. [34] propose a Zero SNR strategy that reduces color discrepancies but requires diffusion models trained with v-prediction objective, which SD1.5 is not. Beneficially, Stable Diffusion v2.1 employs a v-prediction objective. Therefore we replace SD1.5 with SD2.1 while maintaining our previously described learning protocol.

We call this multi-step variation with augmented SD2.1 UNet AUGUNET. AUGUNET solves the misalignment issue and reduces the color shift significantly as shown in Figure 8, resulting in the generation of high-quality, sharp scene intrinsics with much improved quantitative accuracy. In Figure 9 , we show how AUGUNET performs on unseen 1024^2 synthetic domain while trained exclusively on 512^2 real-world images. However, quantitatively, the results still fall short of our SD Unet (single step) result, DINOv2 and Omnidatav2. In the future, we hope this problem will be solved by improved sampling techniques and the next generation of generative image models.

6. Conclusion

We find consistent, compelling evidence that generative models implicitly learn physical scene intrinsics, allowing tiny LoRA adaptors to extract this information with minimal fine-tuning on labeled data. More powerful generative models produce more accurate scene intrinsics, strengthening our hypothesis that learning this information is a natural byproduct of learning to generate images well. Finally, across various generative models and the self-supervised DINOv2, scene intrinsics exist in their encodings resonating with fundamental "scene characteristics" as defined by Barrow and Tenenbaum [3].

We hope that future work expands on these findings. For example explicitly incorporating the production of scene intrinsics into the learning process of generative image models, or developing evaluation metrics for generative models based on physical properties.

Acknowledgments

We would like to thank Shenlong Wang for his insights on LoRA.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Labels4free: Unsupervised segmentation using stylegan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13970–13979, 2021. 4

[2] Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Generative modeling for multi-task visual learning. In *International Conference on Machine Learning*, pages 1537–1554. PMLR, 2022. 4

[3] H Barrow and J Tenenbaum. Recovering intrinsic scene characteristics. *Comput. vis. syst*, 2(3-26):2, 1978. 4, 9

[4] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020. 3

[5] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3, 4, 5, 6, 7

[6] Anand Bhattad and D.A. Forsyth. Stylitgan: Prompting stylegan to generate new illumination conditions. In *arXiv*, 2023. 4

[7] Anand Bhattad and David A Forsyth. Cut-and-paste object insertion by enabling deep image prior for reshading. In *2022 International Conference on 3D Vision (3DV)*, pages 332–341. IEEE, 2022. 3, 5, 6, 4

[8] Anand Bhattad, Daniel McKee, Derek Hoiem, and DA Forsyth. Stylegan knows normal, depth, albedo, and more. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3, 4

[9] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 9

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 4

[11] Yida Chen, Fernanda Viégas, and Martin Wattenberg. Beyond surface statistics: Scene representations in a latent diffusion model. *arXiv preprint arXiv:2306.05720*, 2023. 2, 4

[12] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 4

[13] Katherine Deck and Tobias Bischoff. Easing color shifts in score-based diffusion models. *arXiv preprint arXiv:2306.15832*, 2023. 9

[14] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 4, 5

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 4

[16] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. 2, 4, 6

[17] David Forsyth and Jason J Rock. Intrinsic image decomposition using paradigms. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7624–7637, 2021. 4, 5

[18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 3

[19] Yulu Gan, Sungwoo Park, Alexander Schubert, Anthony Philippakis, and Ahmed Alaa. Instructcv: Instruction-tuned text-to-image diffusion models as vision generalists. *arXiv preprint arXiv:2310.00390*, 2023. 4

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[21] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 3

[22] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *arXiv preprint arXiv:2305.15581*, 2023. 2, 4

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2, 4

[26] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021. 4

[27] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[28] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18974, 2022. 3, 4, 5, 6, 7, 9, 1, 2

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2, 3

[30] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020. 6

[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2, 3

[32] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 2, 4

[33] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8300–8311, 2021. 4

[34] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. *arXiv preprint arXiv:2305.08891*, 2023. 9

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 4

[36] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 1

[37] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, 2023. 2, 4

[38] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1

[39] Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via visual prompting. *arXiv preprint arXiv:2307.14331*, 2023. 1, 3

[40] Atsuhiro Noguchi and Tatsuya Harada. Rgbd-gan: Unsupervised 3d representation learning from natural image datasets via rgbd image synthesis. In *International Conference on Learning Representations*, 2020. 4

[41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4, 7, 9

[42] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. In *International Conference on Learning Representations*, 2021. 4

[43] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[44] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 4

[45] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 4

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 4

[47] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora (https://github.com/cloneofsimo/lora 4

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[49] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR 2023–IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 4

[50] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 2

[51] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 2, 4

[52] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 4

[53] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. 4

[54] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 7

[55] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 3

[56] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. *arXiv preprint arXiv:2303.04803*, 2023. 4

[57] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. 2

[58] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multimodal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023. 2

[59] Ning Yu, Guilin Liu, Aysegul Dundar, Andrew Tao, Bryan Catanzaro, Larry S Davis, and Mario Fritz. Dual contrastive loss and attention for gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6731–6742, 2021. 2

[60] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. What does stable diffusion know about the 3d scene? *arXiv preprint arXiv:2310.06836*, 2023. 2, 4

[61] Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. In *International Conference on Learning Representations*, 2021. 4

[62] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021. 4

[63] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *ICCV*, 2023. 2, 4

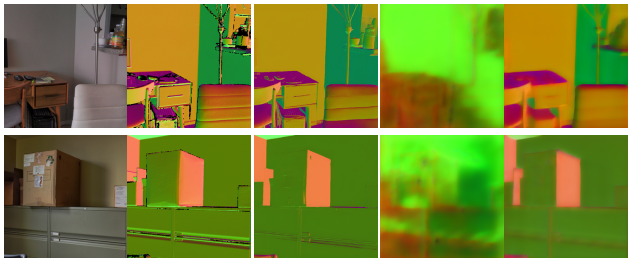
Generative Models: What do they know? Do they know things? Let's find out!

Supplementary Material

7. Control Experiment

To understand whether the intrinsic images extracted by our approach are primarily attributed to the LoRA layers or if they emerge from the underlying generative pretraining of the models themselves, we conducted a control experiment. We adopted the same single-step training protocol used with our SD UNet model, but with a crucial modification: instead of SDv1.5, we used a UNet with random initial weights. The results, as shown in Figure 12, clearly indicate the randomly initialized model's bad performance, both in numerical metrics and visual quality. This suggests that the ability to extract surface normals is not merely a byproduct of our I-LoRA layers but significantly relies on the sophisticated feature representations developed during the generative pretraining of the model.

Mean Angular Error°↓	18.90	36.18	20.31
L1 Error (× 100) ↓	15.21	29.28	16.53



(a) Real (b) GT (c) Omni-v2 [28] (d) Random (d) SD v1.5

Figure 12. Control experiment to test the effectiveness of using a randomly initialized UNet when compared to Stable Diffusion pretrained UNet for extracting surface normals using I-LoRA. The results of this experiment are presented in column (d), while the results of our SD UNet (single step) model are presented in column (e). Both experiments were identical except for the pretrained weights used in the UNet. A UNet that has been initialized randomly has difficulties in extracting surface normal-like representations. Based on our findings, we can conclude that the intrinsic representation is an emergent property of generative pretraining.

8. Additional Ablation Studies

8.1. Number of Diffusion Steps

To assess the impact of the number of diffusion steps on the performance of multi-step AUGUNET models, we conducted an ablation study. The results are presented in Figure 13. For all our experiments in the main text, we used DPMSolver++ [36]. Interestingly, the quality of results did not vary significantly with an increased number of steps, indicating that 10 steps are sufficient for extracting better surface normals from the Stable Diffusion. Nevertheless, we use 25 steps for all our experiments because it is more stable across different image intrinsics.

8.2. CFG scales

When using multi-step AUGUNET , the quality of the final output is influenced by the choice of classifier-free guidance (CFG) during the inference process. In Figure 14, we present a comparison of the effects of using different CFG scales. Based on our experiments, we found that using CFG=3.0 results in the best overall quality and minimizes color-shift artifacts.

9. Extracting Intrinsics with other Stable Diffusion Image Editing Methods

We experiment with other Stable diffusion-based image editing methods that can learn certain concepts/styles without the need to finetune the entire network. In Figure 15, we show that without the additional image latent encoding we proposed as AUGUNET, SDEdit [38] has difficulty in generating aligned surface normals for the given image. Moreover, our attempts to integrate a distinct style token for "surface normal" using the Textual Inversion technique [18], as well as optimizing token embeddings in the manner of VISII [39], did not yield satisfactory outcomes, as evidenced in Figure 16.

10. Hyper-parameters

In Table 4, we show the hyperparameters we use for each model.

11. Additional Qualitative Results

In Figure 17, we present more results for AUGUNET and AUGUNET1.5 . Figure 18 shows extra results for models trained on FFHQ dataset. More examples of scene intrinsics extracted from StyleGAN-v2 trained on LSUN bedroom can be found in Figure 19. In Figure 20, we show results for SD Unet (single step) on generated images. Shown in Figure 21 are extra results for StyleGAN-XL trained on ImageNet.

12. Results on 1024^2 synthetic images

Our AUGUNET models, although trained exclusively on 512^2 images from the DIODE dataset, demonstrate their robustness by successfully extracting intrinsic images from 1024^2 high-resolution synthetic images generated by Stable Diffusion XL [43], as shown across Figures 22 to 31

Mean Angular Error°↓	25.83	23.79	23.48	23.86	23.79	23.74	23.67
L1 Error (× 100) ↓	21.08	19.39	19.10	19.40	19.35	19.31	19.25

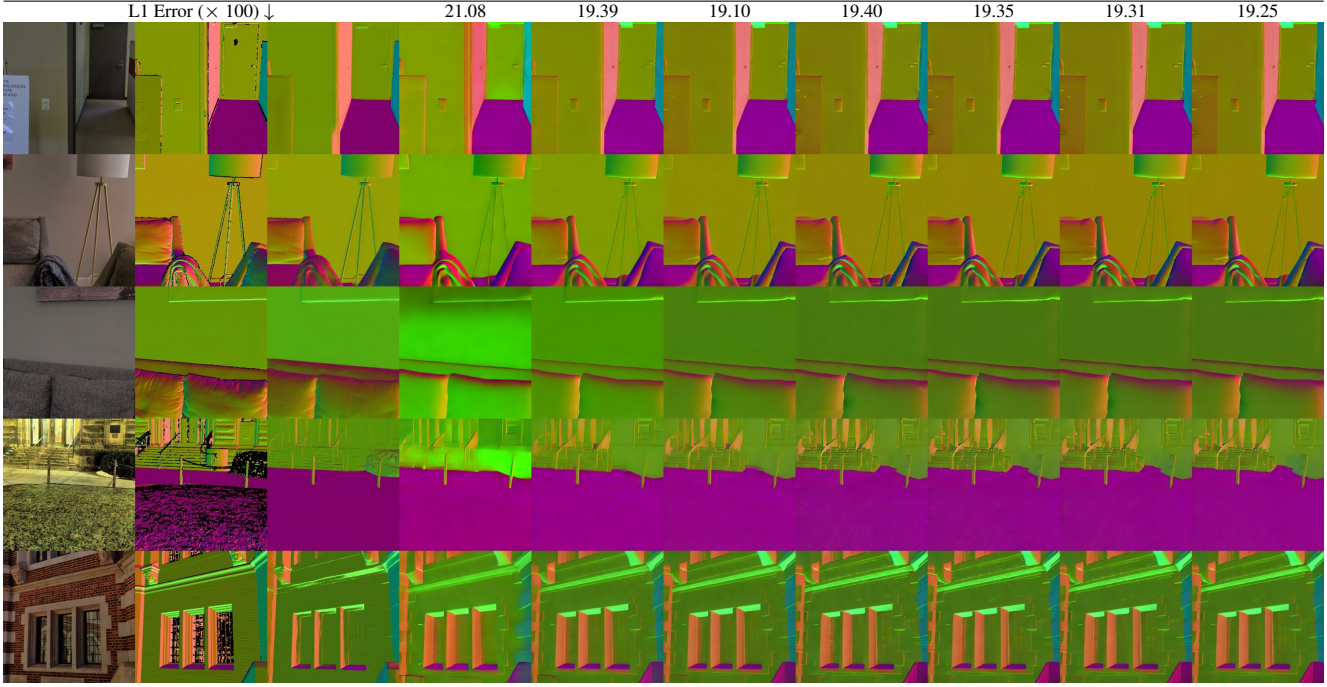


Image | GT | Omni-v2 [28] | Steps=2 | Steps=5 | Steps=10 | Steps=15 | Steps=20 | Steps=25 | Steps=50

Figure 13. Ablation study to determine the effect of varying numbers of diffusion steps while keeping CFG fixed at 3.0. Our findings show that there are very small differences, both in terms of quantity and quality, after 10 steps. For our main paper, we report results for 25 steps as it is more stable across different intrinsics.

Model	Dataset	Resolution	Rank	LR	BS	LoRA Params	Generator Params	Steps Till Convergence
VQGAN	FFHQ	256	8	1e-03	1	0.13M	873.9M	$\sim$ 1500
StyleGAN-v2	FFHQ	256	8	1e-03	1	0.14M	24.8M	$\sim$ 2000
StyleGAN-v2	LSUN Bedroom	256	8	1e-03	1	0.14M	24.8M	$\sim$ 2000
StyleGAN-XL	FFHQ	256	8	1e-03	1	0.19M	67.9M	$\sim$ 1000
StyleGAN-XL	ImageNet	256	8	1e-03	1	0.19M	67.9M	$\sim$ 2500
AUGUNET (multi step)	Open	512	8	1e-04	4	1.59M	943.2M	$\sim$ 30000
SD UNet (single step)	Open	512	8	1e-04	4	1.59M	943.2M	$\sim$ 15000

Table 4. Hyper-parameters for each model. LR refers to the learning rate and BS refers to the batch size. Please note that the number of steps required to reach convergence reported above is for normal/depth. However, it is worth noting that albedo and shading tend to require significantly fewer steps to converge. Additionally, AUGUNET and SD UNet are trained on real-world DIODE dataset, while the other models are trained on synthetic images within a specific domain. (Num. of params of VQGAN counts transformer + first stage models; Num. of params of AUGUNET and SD UNet counts VAE+UNet)

Mean Angular Error°↓	24.28	23.48	25.72	27.80	29.85	31.93	34.12
L1 Error (× 100) ↓	19.48	19.10	21.01	22.72	24.36	26.03	27.78

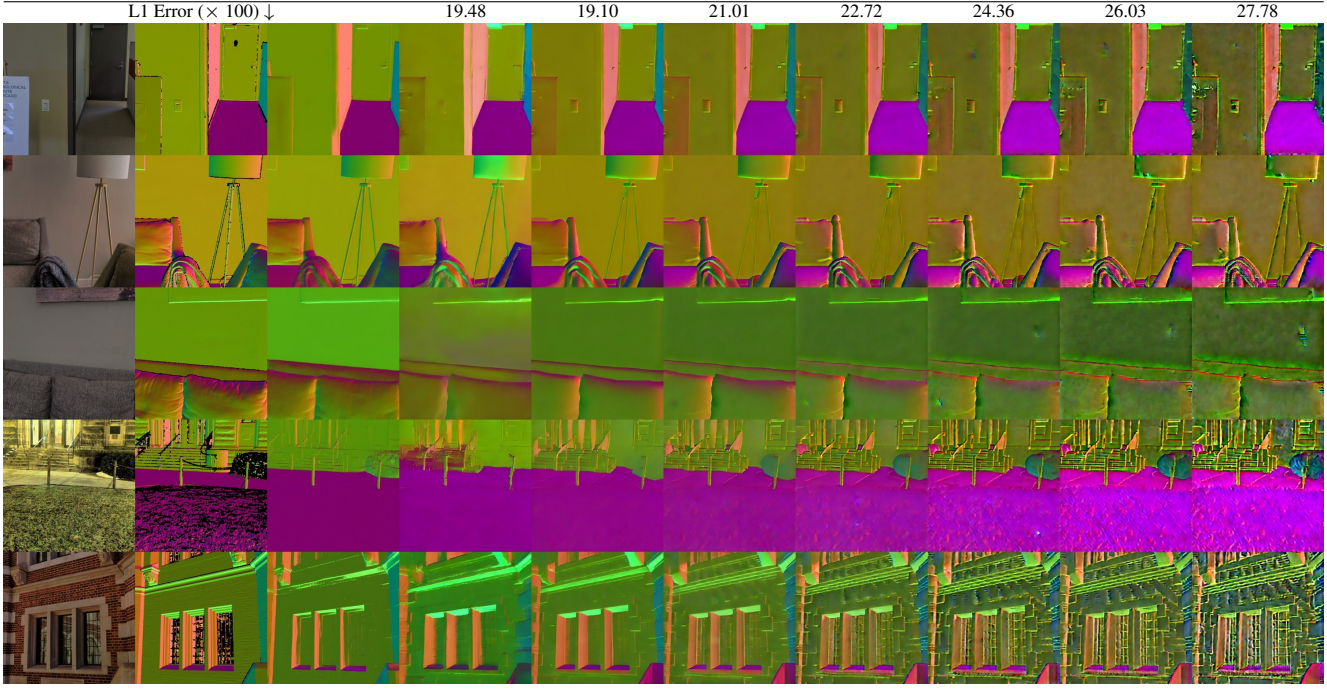


Image GT Omni-v2 [28] CFG=1 CFG=3 CFG=5 CFG=7 CFG=9 CFG=11 CFG=13

Figure 14. Ablation study analyzing the impact of different classifier-free guidance (CFG) on AUGUNET surface normal prediction. For efficiency, we experimented with a step of 10. We observed that CFG=1 sometimes led to incorrect semantic predictions, particularly in the case of stairs in row 4. On the other hand, using large CFGs (5 and beyond) results in severe color shift problems.

Image s=0.2 s=0.5 s=0.7 s=1.0

Figure 15. We observe applying SDEdit method on the SD1.5 model alone, without incorporating the additional input image latent encoding, fails to produce satisfactorily aligned and high-quality scene intrinsics. The reason for this is the considerable domain shift that exists between RGB images and surface normal maps, which results in severe artifacts when using SDEdit. The variable "s" represents the strength of SDEdit.

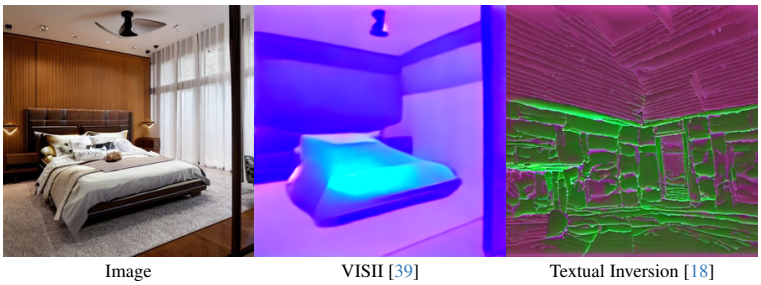
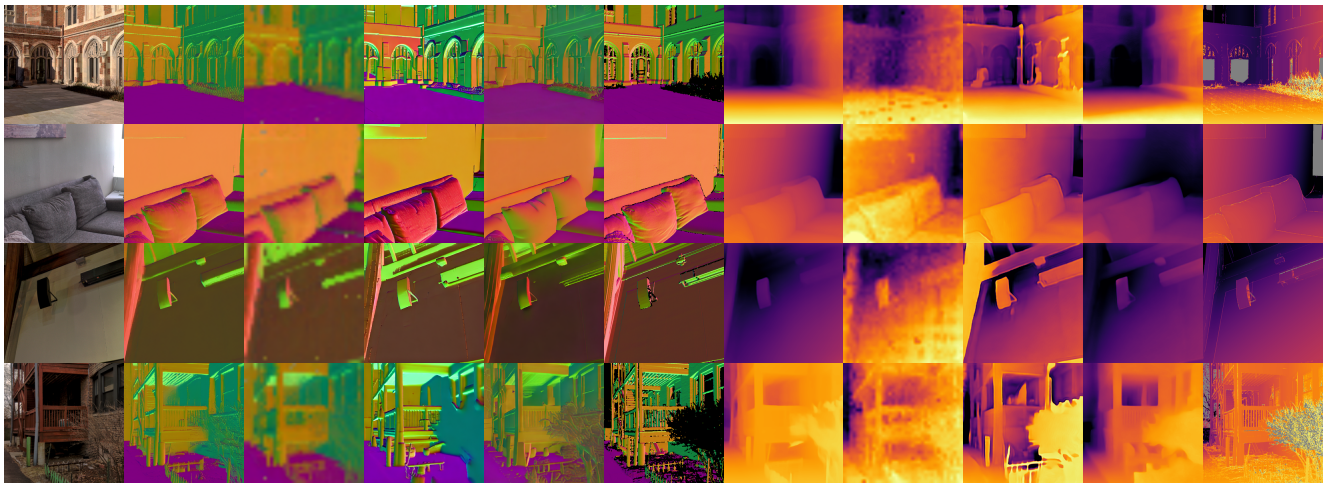


Image VISII [39] Textual Inversion [18]

Figure 16. We observe that applying VISII or Textual Inversion for learning surface normal results in unsatisfactory output. Although VISII can retain some geometrical consistency, it is unable to learn surface normal. On the other hand, Textual Inversion fails completely at the task.



(a) Real Image (b) Omni-v2 (c) DINOv2 (d) AUGUNET1.5 (e) AUGUNET (f) GT (g) ZoeDepth (h)DINOv2 (i) AUGUNET1.5 (j) AUGUNET (k) GT

Figure 17. Additional results after applying improved diffusion techniques with AUGUNET. AUGUNET was found to significantly reduce color shift artifacts observed in AUGUNET1.5 during the extraction of detailed scene intrinsic results.

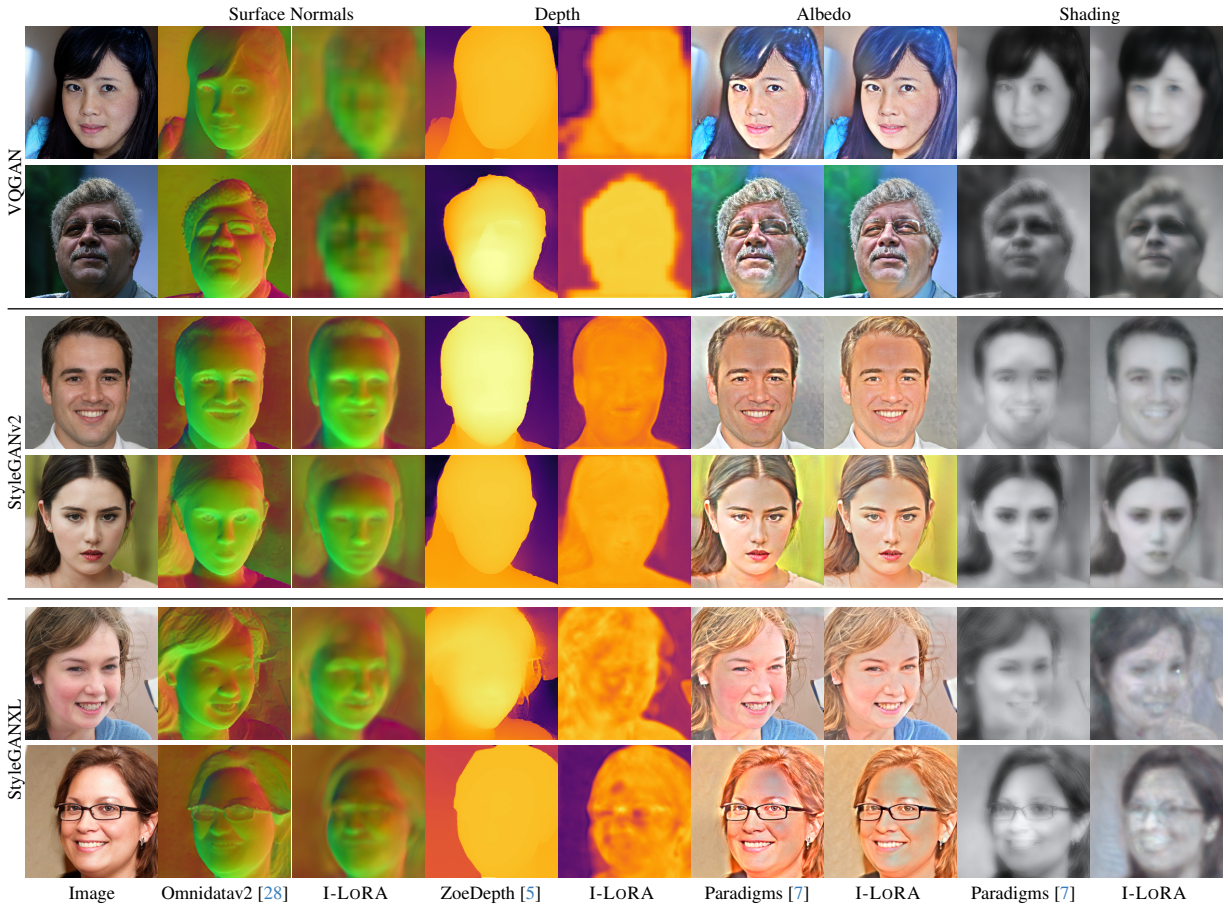


Figure 18. Additional results of scene intrinsics from different generators – VQGAN, StyleGAN-v2, and StyleGAN-XL – trained on FFHQ dataset.

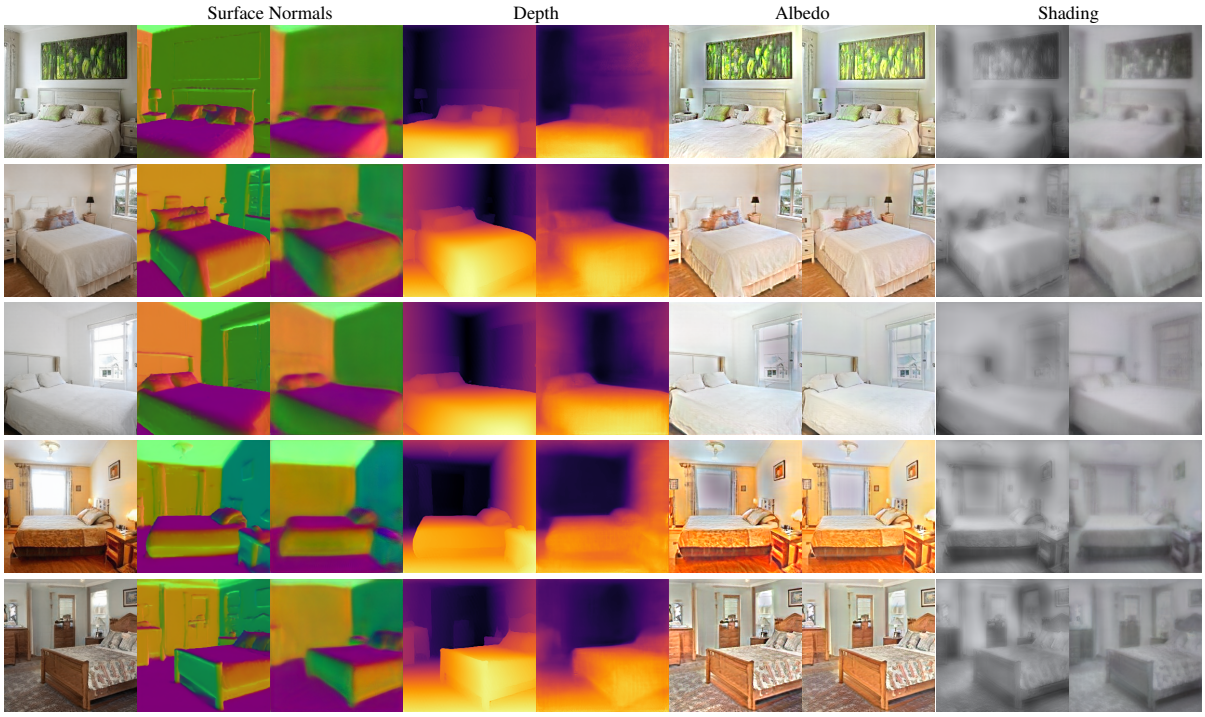


Figure 19. Additional results of scene intrinsics extraction from Stylegan-v2 trained on LSUN bedroom images.

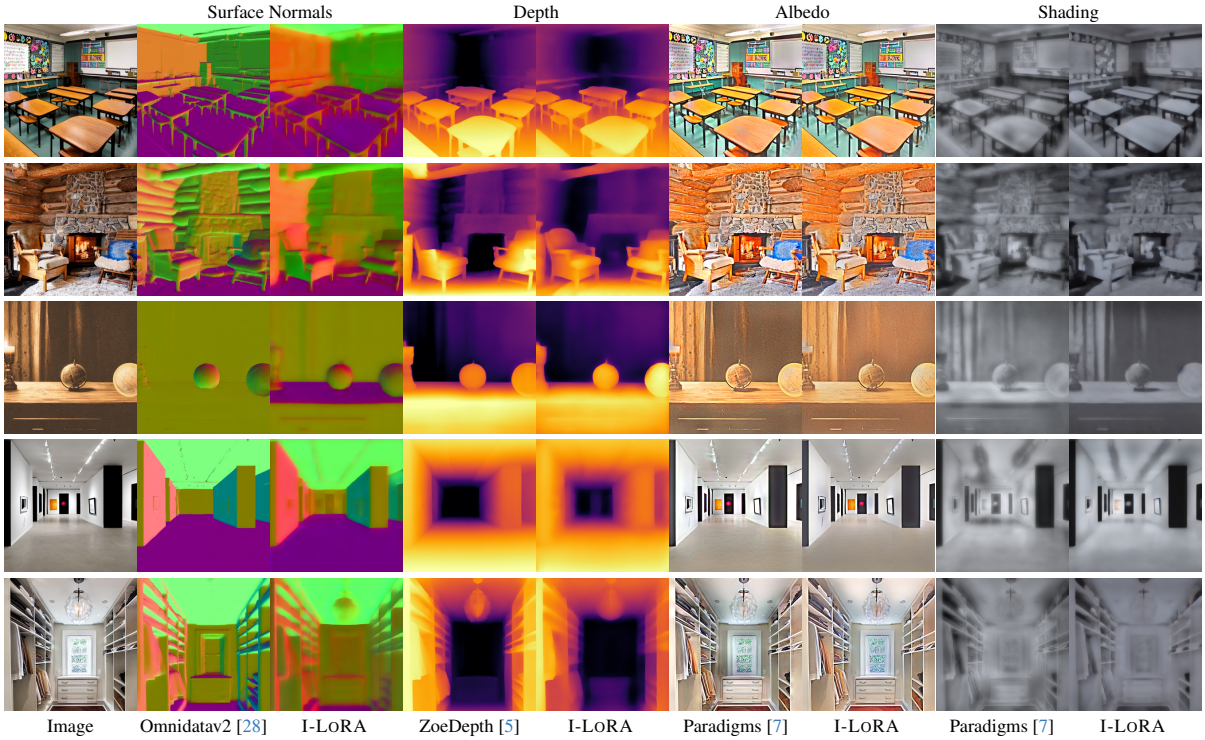


Figure 20. Additional results for SD UNet v1.5 (single step). Note on the third row, our model correctly predicts the surface normal of the table.

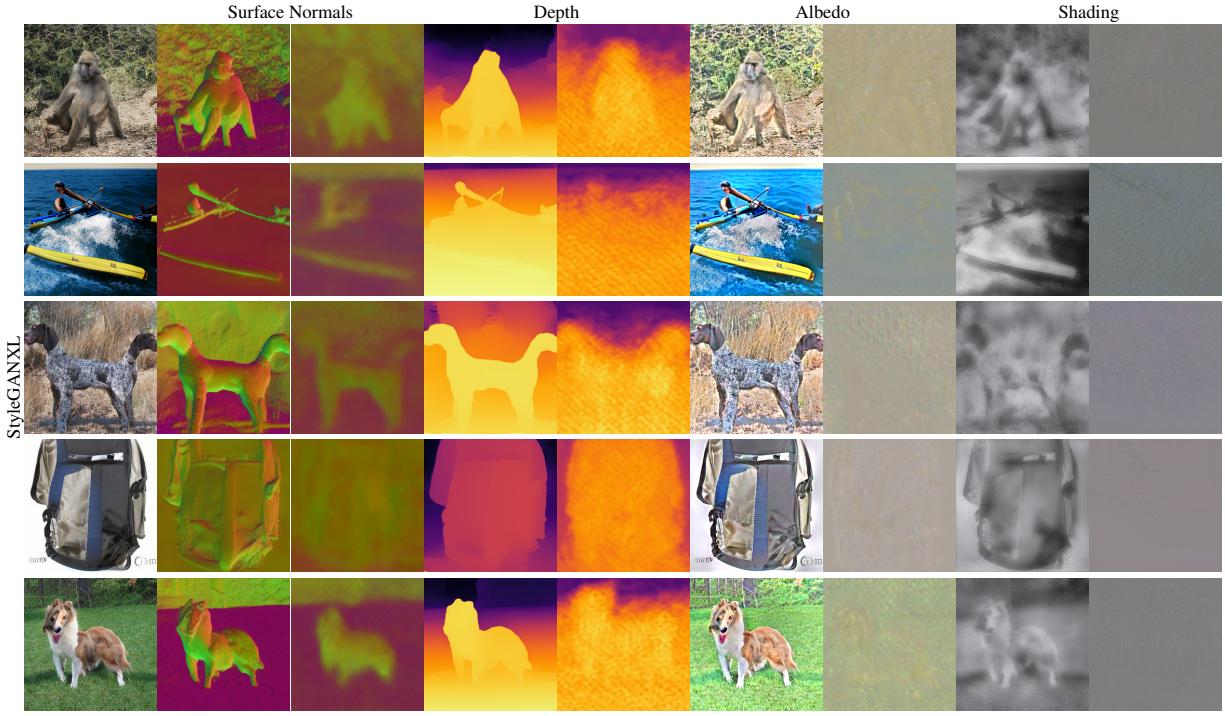


Figure 21. Additional results for StyleGAN-XL trained on ImageNet. StyleGAN-XL's inability to produce image intrinsics may be due to its inability to create high-quality plausible images.

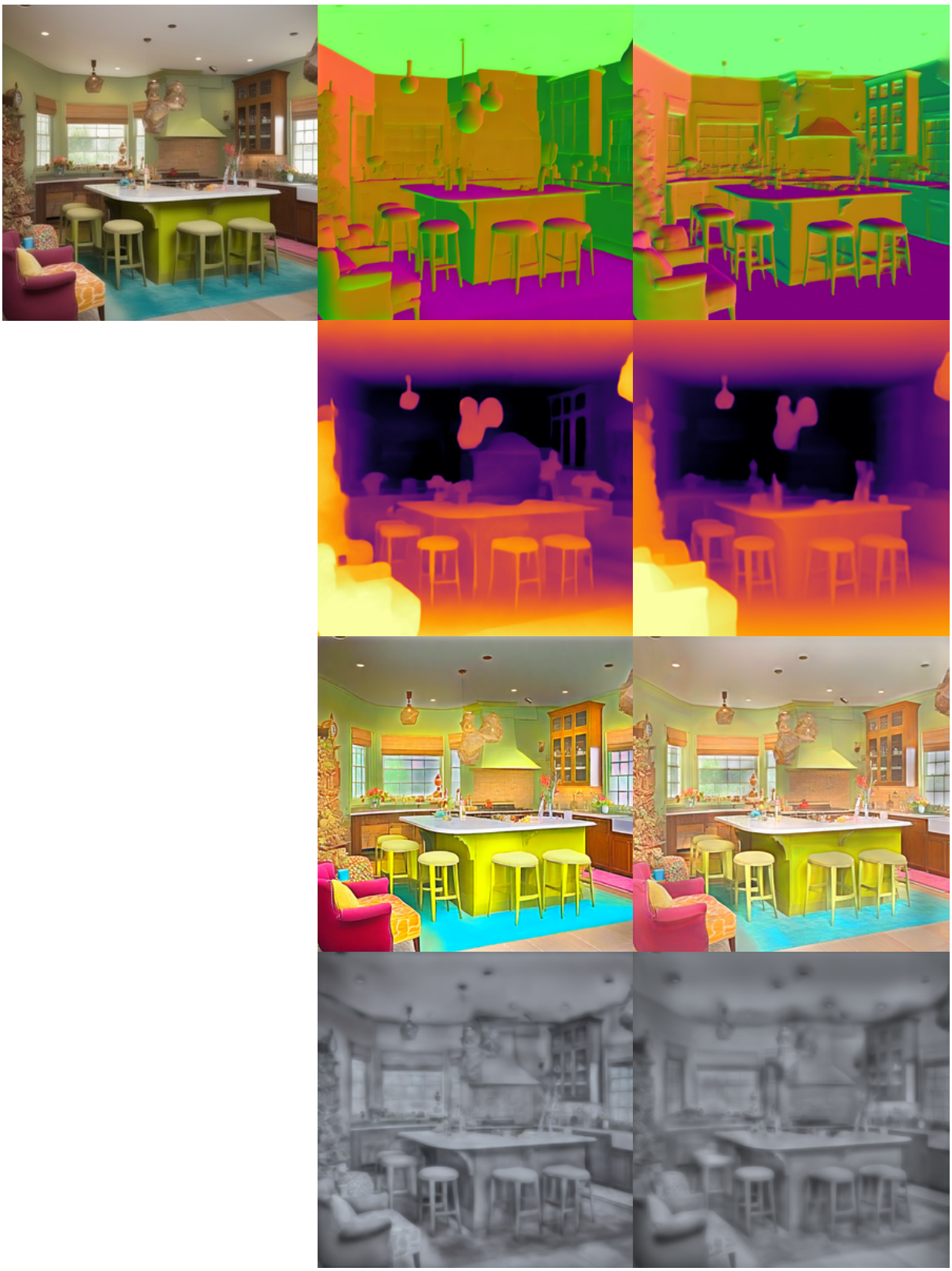


Figure 22. Results of AUGUNET models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

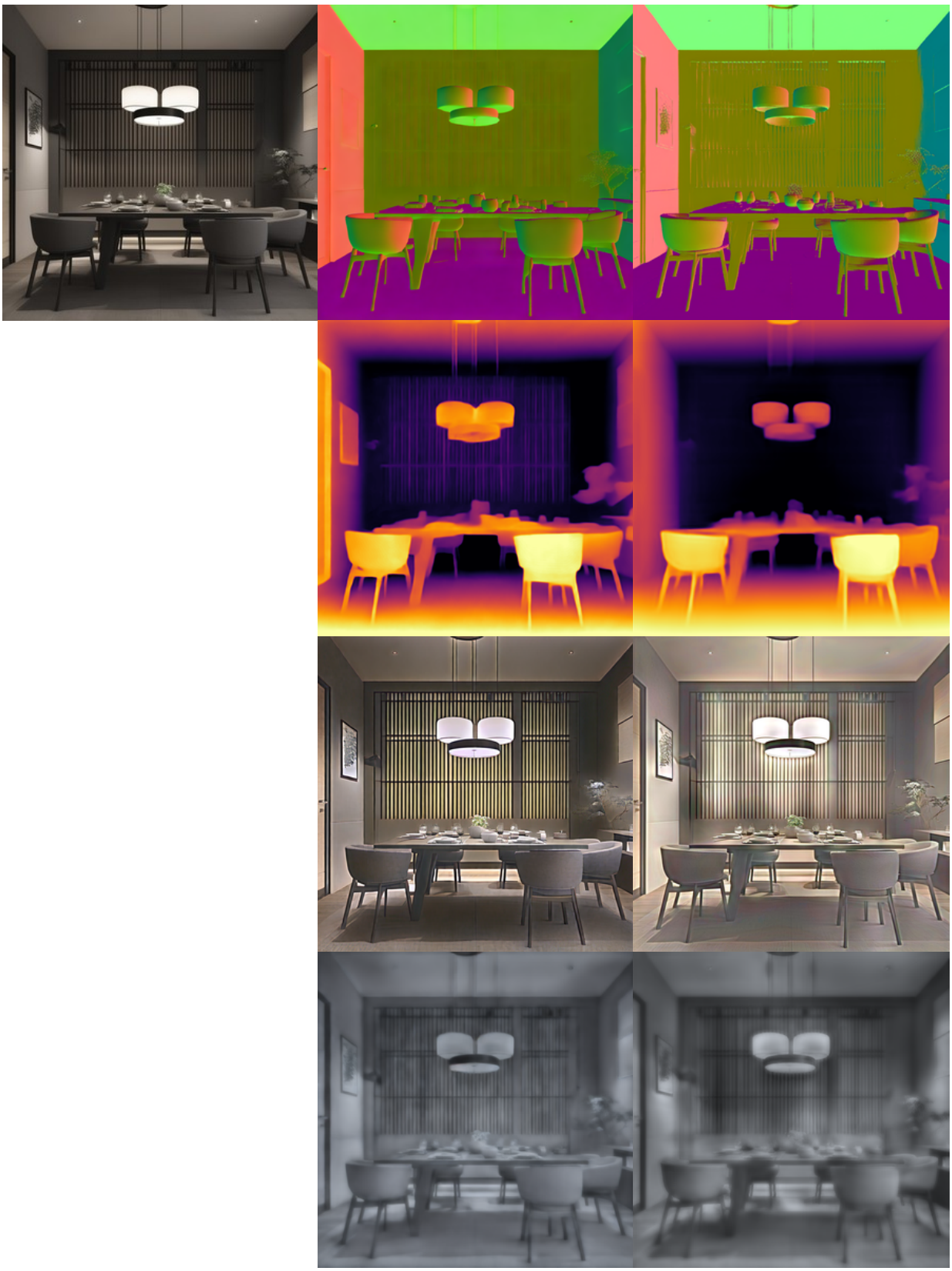


Figure 23. Cont. results of AUGUNET models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

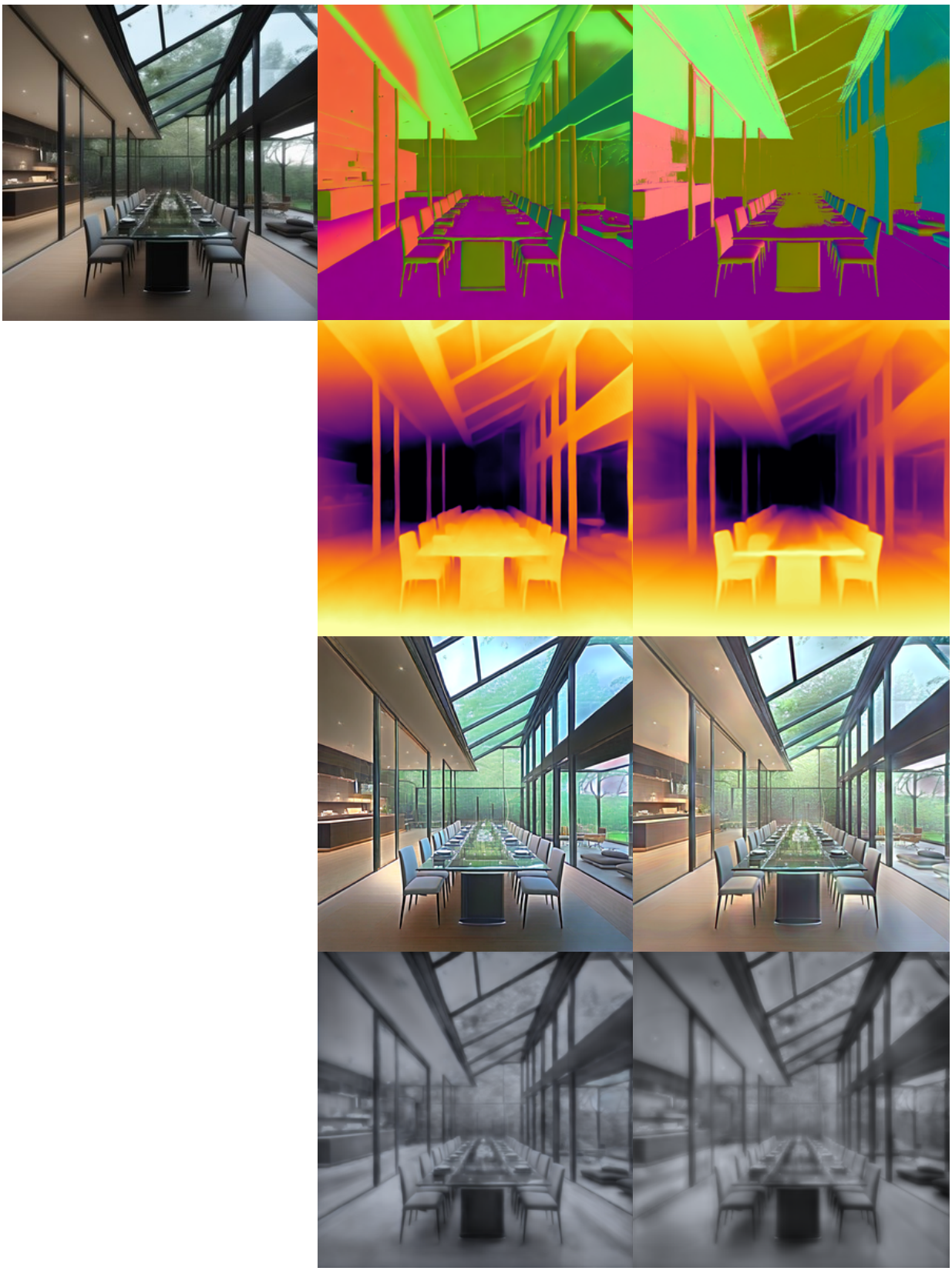


Figure 24. Cont. results of AUGUNET models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

Figure 25. Cont. results of AUGUNET models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

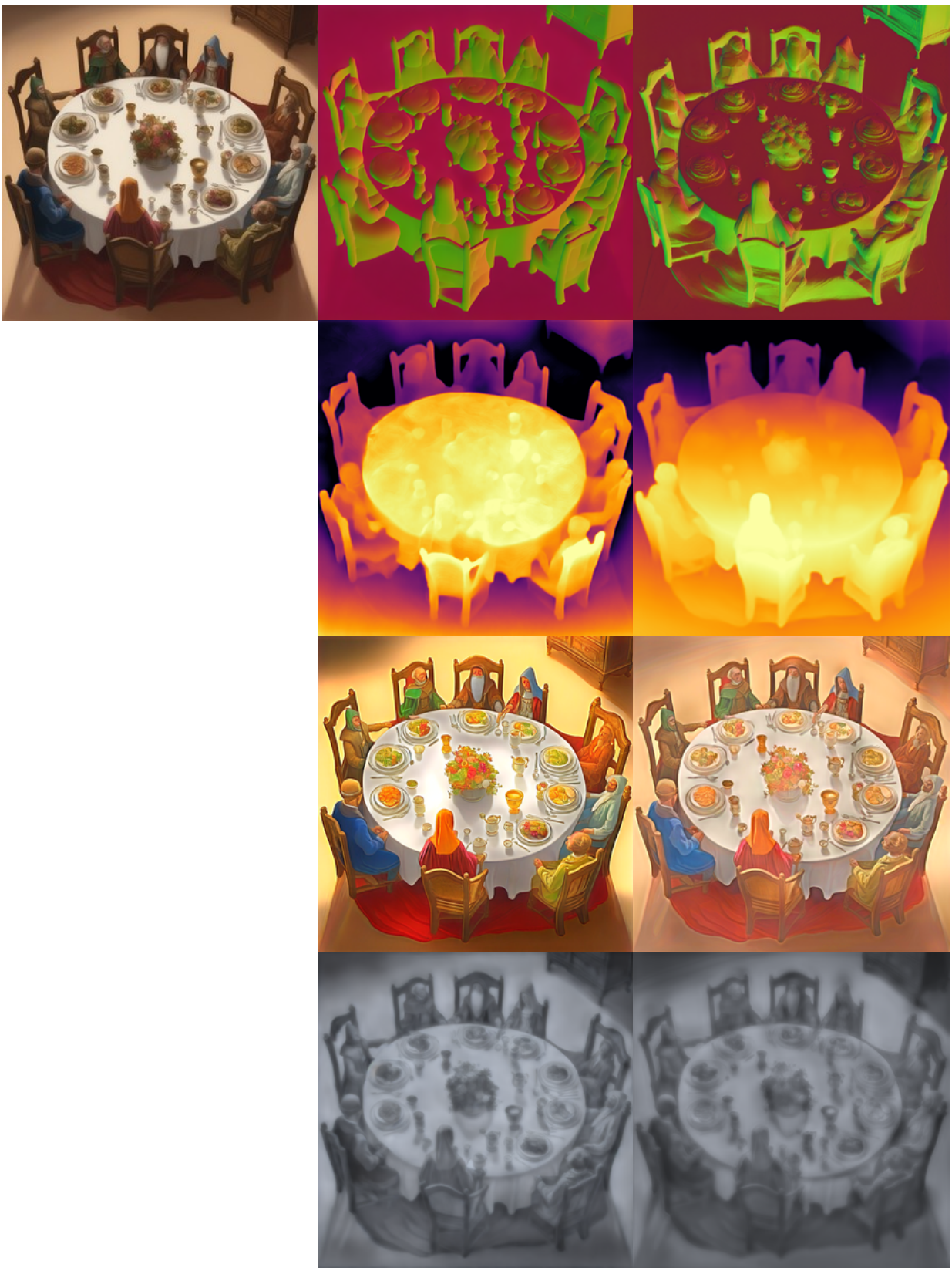


Figure 26. Cont. results of AUGUNET models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

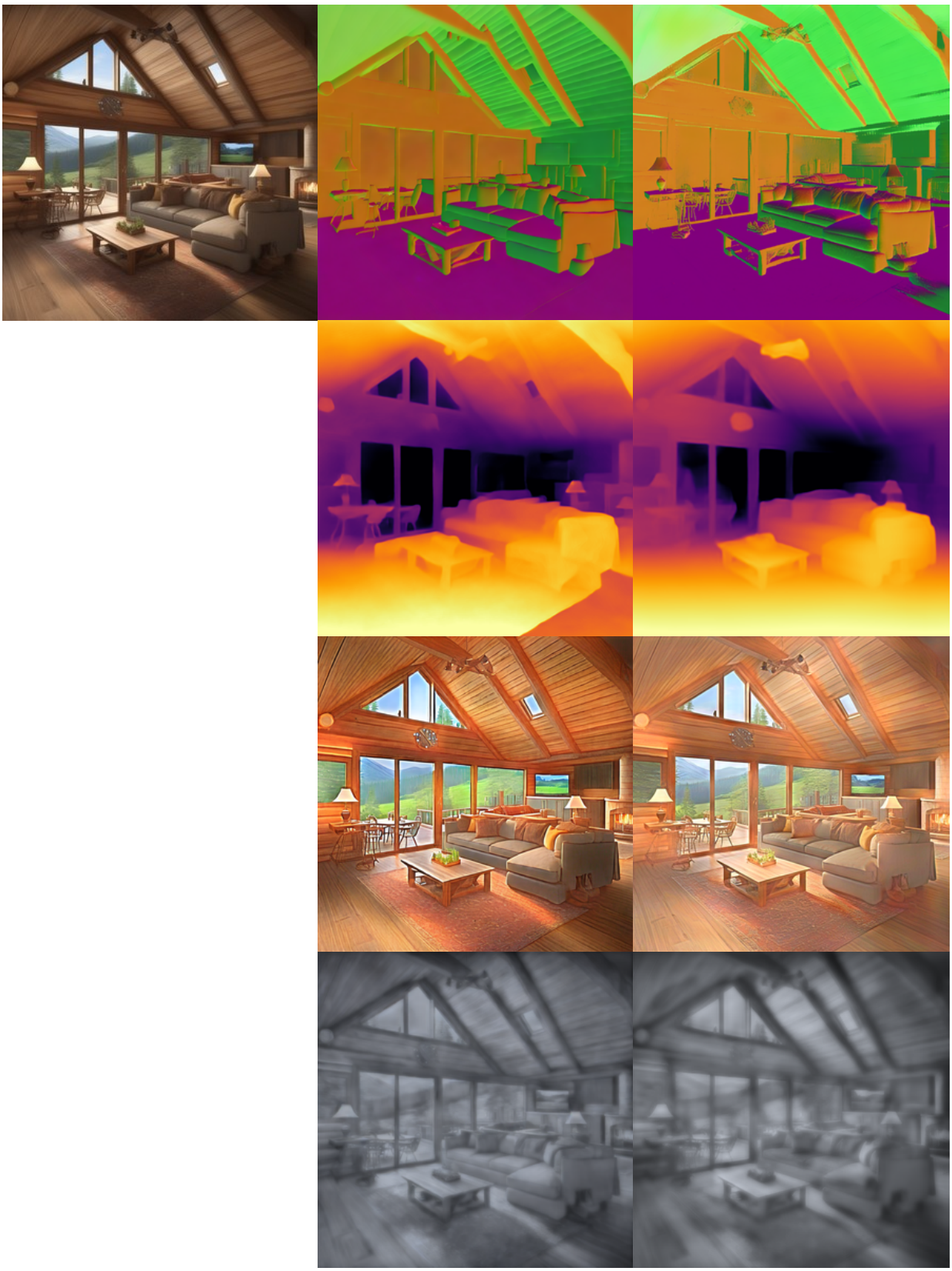


Figure 27. Cont. results of AUGUNET models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

Figure 28. Cont. results of AUGUNET models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

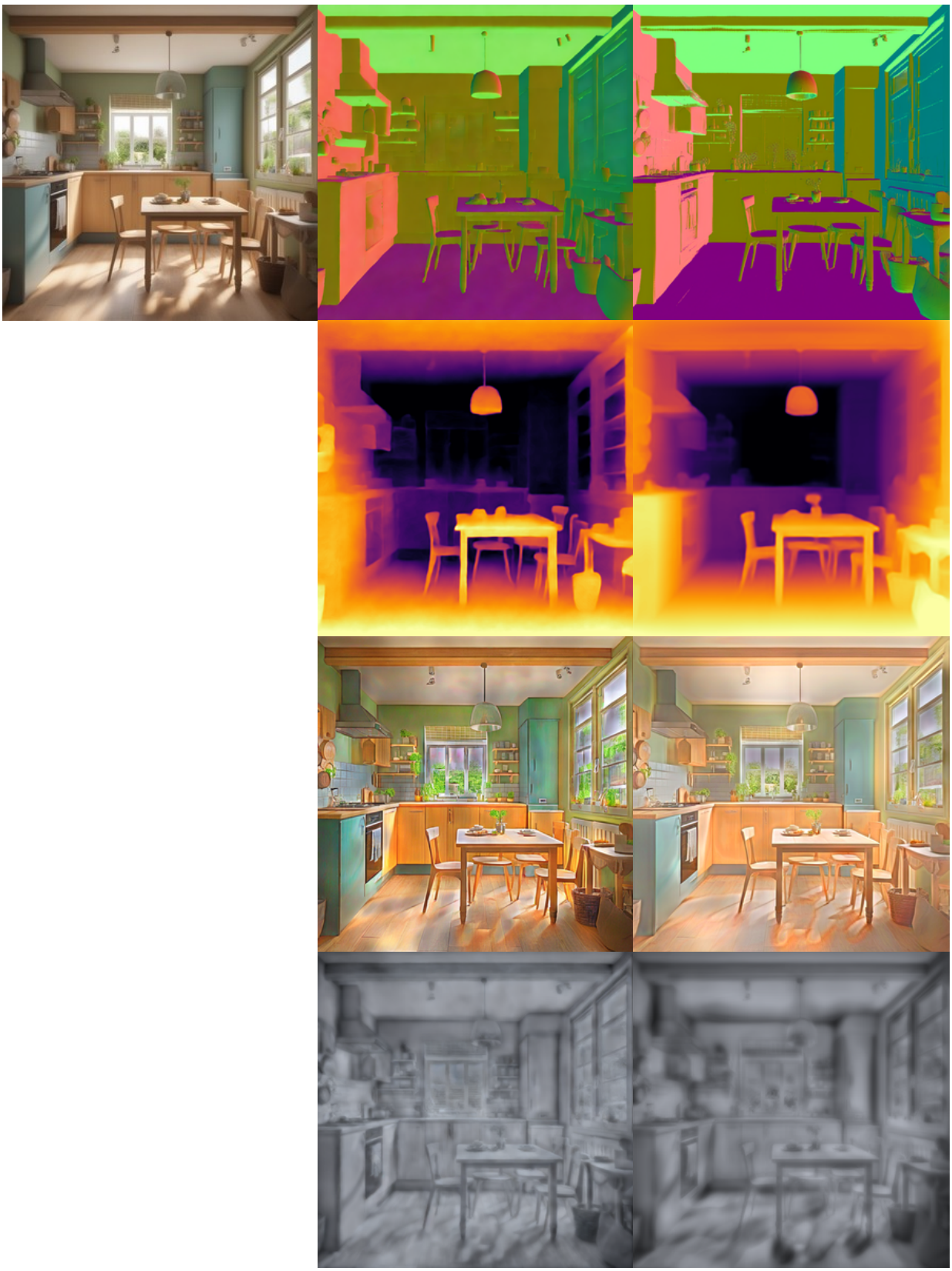


Figure 29. Cont. results of AUGUNET models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

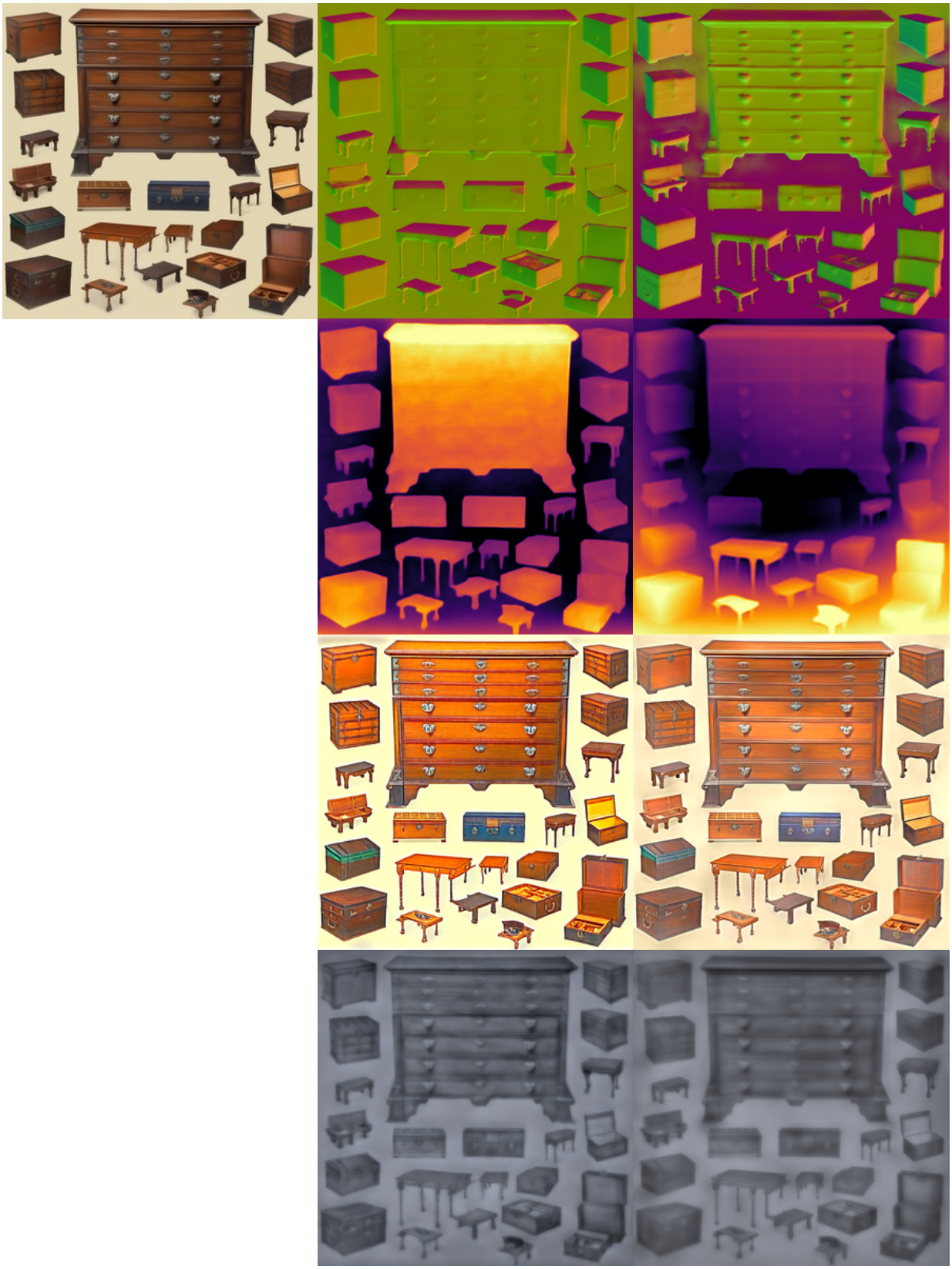


Figure 30. Cont. results of AUGUNET models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

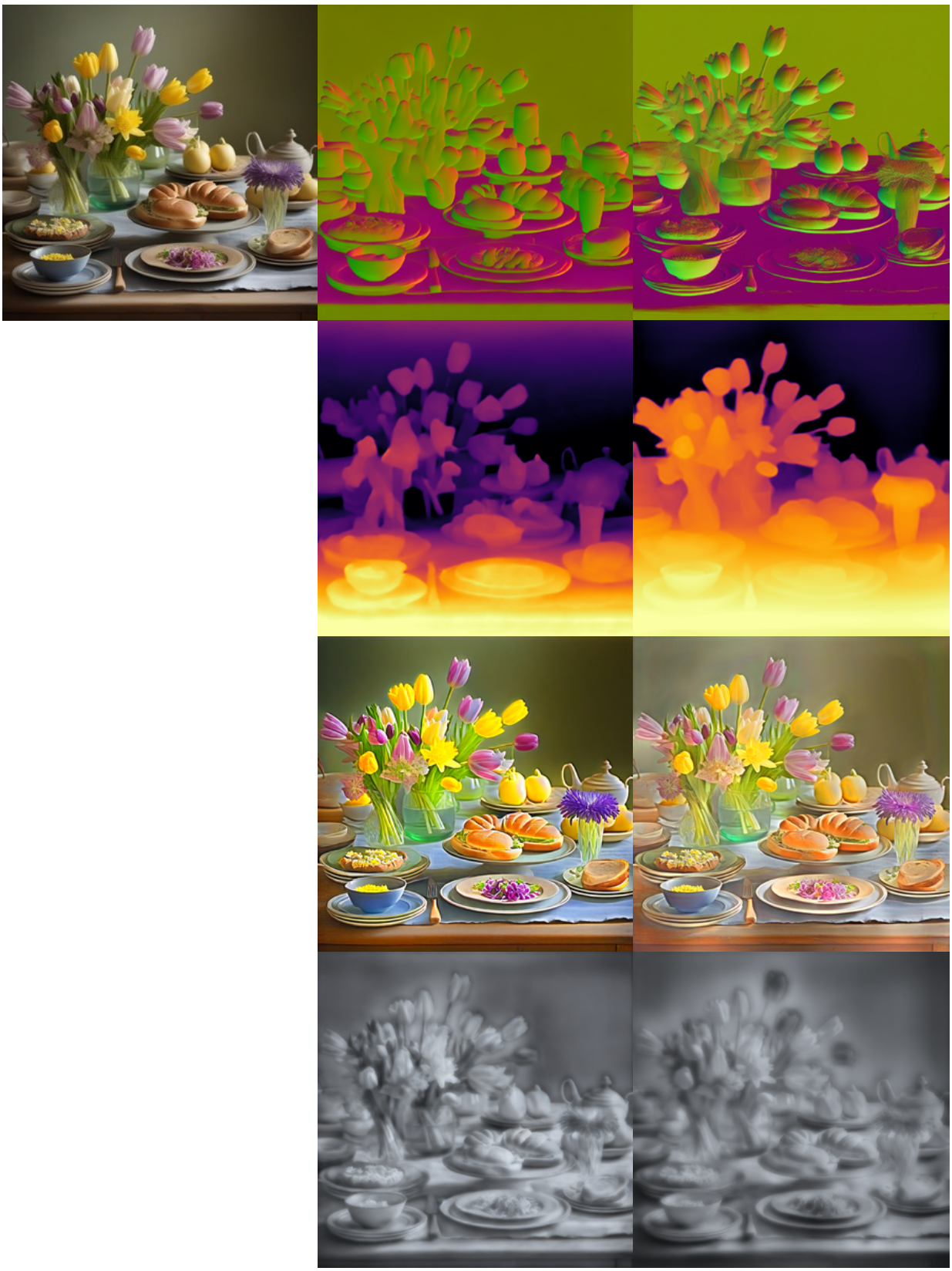


Figure 31. Cont. results of AUGUNET models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.