Diffusion Model Scene Representation

Atharva Kulkarni apkulkarni@ucsd.edu
Ester Tsai etsai@ucsd.edu
Karina Chen kac009@ucsd.edu
Zelong Wang zew013@ucsd.edu
Alex Cloninger acloninger@ucsd.edu
Rayan Saab rsaab@ucsd.edu

Abstract

Many AI image generation tools use diffusion models to create images from text prompts, including Stable Diffusion, which uses a latent diffusion model (LDM). Even though these LDMs are trained on 2D images, the resulting images contain detailed, coherent 3D scenes. This research explains and visualizes how the LDM encodes depth and shading. By probing the internal activations of the LDM's neural network using linear probing classifiers, we conclude that 3D information like depth, saliency, and shading are already encoded by step 2 of the denoising process. This information is present early in the denoising process before our human eyes can detect anything but noise. This has implications for improving the diffusion process by speeding up training and denoising. In addition, by understanding when information is being encoded in the denoising process, we can prevent potential adversarial attacks.

Website: https://abc.github.io/
Code:
https://github.com/karinaechen/diffusion-model-internal-representation

1 Introduction

The remarkable advancements of Latent Diffusion Models (LDMs) enable the generation of realistic images from textual descriptions. LDMs generate images that contain coherent 3D scene and shading representations, even when trained solely on images lacking explicit depth or shading information. Our project investigates how the LDMs encode depth and shading information in their internal activations. This exploration is pivotal for understanding AI's interpretive capabilities and advancing image synthesis.

We successfully demonstrate that the internal representation of scene geometry is captured much earlier than the human eye can recognize during the diffusion process. As seen in Figure 1, each row starting with "Probe" starts to display relevant information much earlier in the diffusion process than its corresponding non-probe row. Each pair of rows (Mask, Depth, Shading) uses an algorithm to visualize either salient object detection, depth map, or shading map respectively. The top row in each pair of rows is the output of linear probes trained on the image's internal representation (a Tensor), while the bottom row is the output from an off-the-shelf model with the intermediate diffusion images as inputs. This finding concludes that oject detection, depth map, and shading qualities are being encoded in the internal activations of the LDMs far earlier than the human eye can see in the intermediate diffusion images.

Additionally, we utilize VGG-16, a 16-layer-deep convolutional neural network trained on the ImageNet database, to explore when an image classification model would recognize an image. We found that the model is not able to detect the diffused image any earlier than the human eye, and in some cases, is slower.

Implications of this finding include ways to fine-tune the diffusion training process. If most of the information is being encoded in the first 80% of time-steps, we could potentially speed up the remaining 20% of time-steps. Furthermore, by understanding when information is being encoded in the denoising process, we can prevent potential adversarial attacks on the LDM.

1.1 Literature Review

Previous work has attempted to answer this question and has found that there is depth information that emerges in early denoising. Baranchuk et al. (2021) extrapolated the intermediate activations of a pre-trained diffusion model for semantic segmentation. Their high segmentation performance reveals that the diffusion model encodes the rich semantic representations during training for generative tasks. Our work shows that the internal representation of LDM also captures the geometric properties of its synthesized images.

Our paper is heavily derived from the work done by Chen, Viégas and Wattenberg (2023), which found that linear representations of depth and saliency is indeed encoded within the internal activations of the LDM and appears early on in the denoising process. However, we also investigate the encoding of shading information, and show that this is similarly represented as depth and saliency.

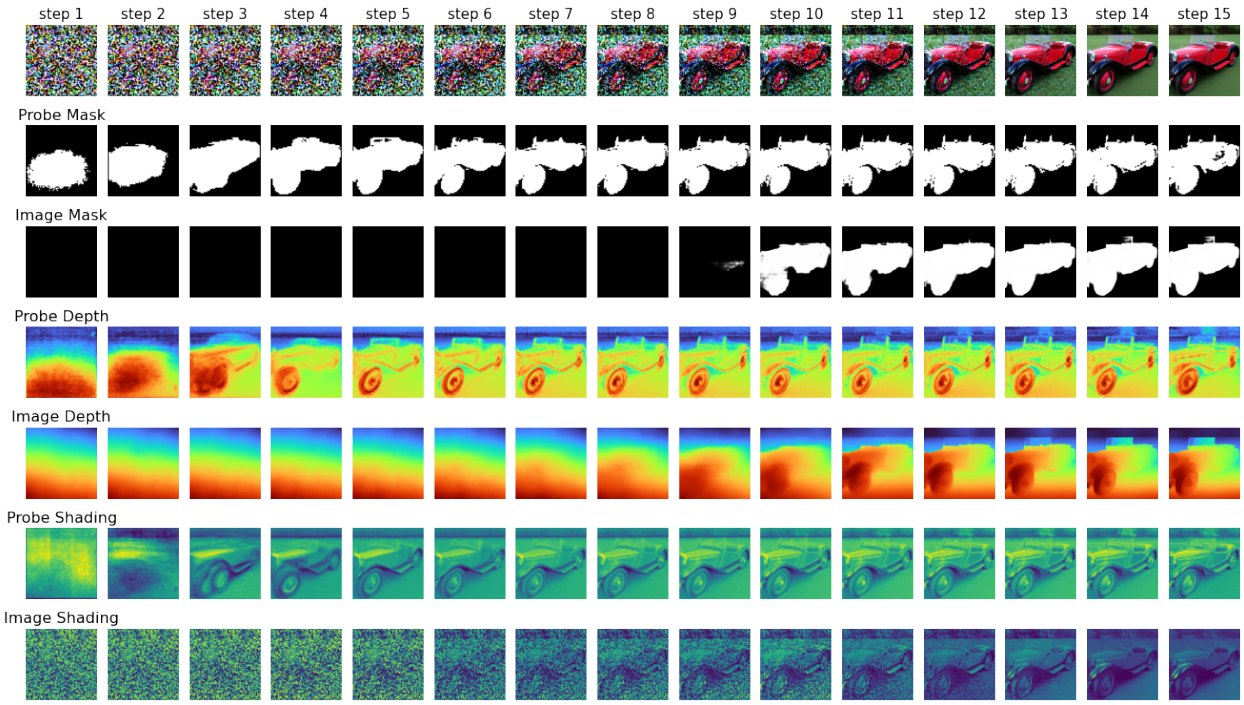

Figure 1: Internal Scene Representations

1.2 Data Description

1.2.1 Generating the Image Dataset

Stable Diffusion is an open-source diffusion model that generates images from text prompts. Stable Diffusion is a two-stage framework that consists of an LDM and (2) a variational autoencoder (VAE). The LDM learns to predict and remove noise by reversing a forward diffusion process. The VAE converts data between latent and image space. After the LDM synthesizes a denoised latent z, the decoder of VAE converts the denoised latent z to the image space.

Starting at V1.5, the Stable Diffusion model has a built-in depth model that can predict a depth map. We generate training images using the older Stable Diffusion v1.4 because it was trained without explicit depth information. Our result shows that even if the LDM is trained without explicit depth information, its internal representation of 3D information still exhibits an impressive ability to capture an underlying model of scene geometry.

Our diffusion image dataset consists of 617 images (512 pixels x 512 pixels) generated using Stable Diffusion v1.4. We have a CSV file that contains the prompt index, text prompt, and seed for each image. For example, the image with the prompt index 5246271, the text prompt ”ZIGGY - EASY ARMCHAIR”, and the seed 64140790 generated the 512 by 512 in Figure 2.

1.2.2 Generating the Ground Truth Images

The diffusion images we generate using Stable Diffusion v1.4 do not have ground truth labels for salient object detection, depth, or shading, so we apply off-the-shelf models to those images to synthesize the ground truth images. The labels are the same size as the diffusion images.

Figure 2: (top left) 512 x 512 image generated by Stable Diffusion v1.4 using the text prompt "ZIGGY - EASY ARMCHAIR" and seed 64140790.
(top right) Salient object detection mask generated by TRACER.
(bottom left) Depth map generated by MiDaS.
(bottom right) Shading and illumination map generated by Intrinsic.

For salient object detection, we apply the salient object tracing model TRACER by Lee, Shin and Han (2022) to generate a mask for each image. The masks are black and white, where white indicates the salient object, or foreground, and black indicates the background.

For depth labels, we apply the pre-trained MiDaS model designed by Ranftl et al. (2020) to the diffusion images to estimate their relative inverse depth maps.

For shading labels, we apply the pre-trained Intrinsic model designed by Careaga and Aksoy (2023) to the diffusion images to generate highly accurate intrinsic decompositions and estimate the shading maps.

2 Methods

Our research method draws inspiration from Chen, Viégas and Wattenberg (2023) and moves beyond their focus on depth to explore other image information such as shading and illumination. We also explore at what point the image recognition models like VGG-16 can correctly identify the image subject in the reverse diffusion process.

2.1 Extracting the Internal Representation

Our research aims to explain and visualize the changes in the LDM's latent space as the LDM generates an original image from pure noise. We extract the self-attention layer's intermediate outputs from the Unet denoising block within the LDM at each denoising step. We use a hook to save the features of a Unet module every time it runs. Each features Tensor has a shape of torch.Size([2, 4096, 320]). We select the features from a specific time step, block type, block index, and layer as the input for the linear probing classifier. The name of the module we select from the Unet is "up_blocks.3.attentions.0.transformer_blocks.0.attn1.to_out.0," which stands for the block "up,", block index 3, layer index 0, and layer name "transformer_blocks.0.attn1.to_out.0."

2.2 Developing the Probing Classifier

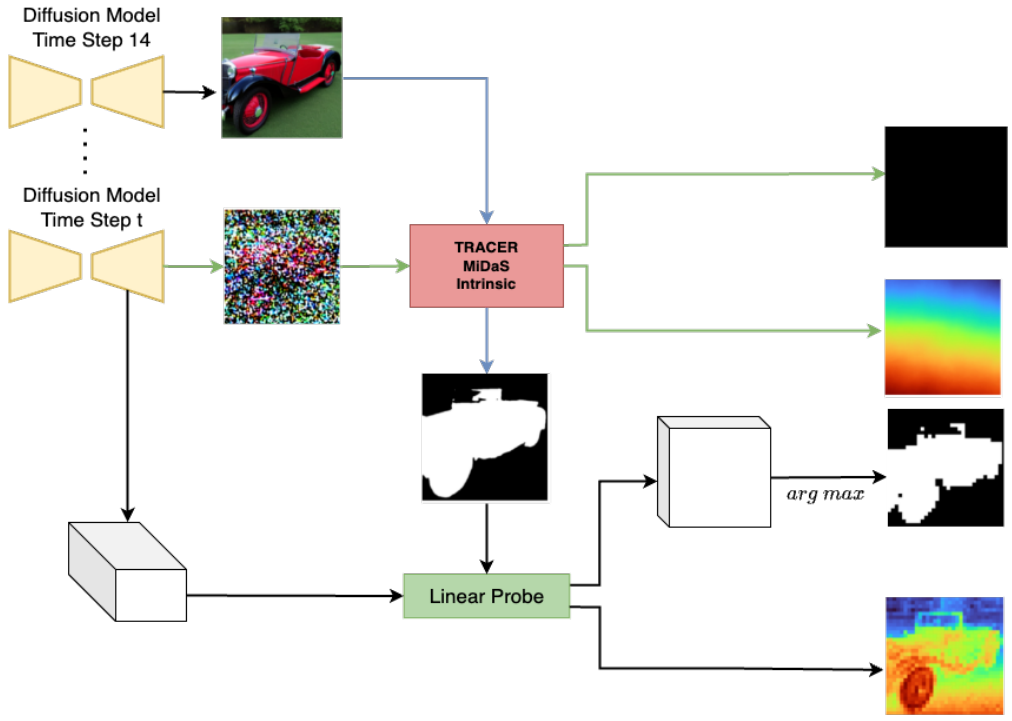

Figure 3: Probing Workflow

The linear probe is a neural network with a linear layer. We train the probe for 30 epochs using the Adam optimizer and cross entropy loss.

The linear probe takes the internal representation of images as its input, with the outputs from the TRACER, MiDaS, and Intrinsic models serving as labels. These labels represent the accurate salient, depth, and shading information derived from the final image.

The probing binary classifier trained on the TRACER image outputs distinguishes foreground from background at the pixel level.

For continuous attributes such as depth and shading, we retrieve the outputs from the self-attention layers and employ a linear regressor trained on these outputs to predict the nuanced variations in depth and shading. The linear regressor trained on the MiDaS image outputs predicts a depth map that shows the 3D scene depicted in the Stable Diffusion image. The linear regressor trained on the Intrinsic image outputs predicts a shading map that shows the illumination depicted in the Stable Diffusion image.

2.3 Choosing the Best Performing Probe

Table 1: Spatial and Feature Dimensions of Self-Attention Layers in the LDM

Blocks	Number of Self-Attn Layers	Spatial $h \times w$	Feature c
Encoder 1	2	64 × 64	320
Encoder 2	2	32 × 32	640
Encoder 3	2	16 × 16	1280
Encoder 4	0	-	-
Bottleneck	1	8 × 8	1280
Decoder 1	0	-	-
Decoder 2	3	16 × 16	1280
Decoder 3	3	32 × 32	640
Decoder 4	3	64 × 64	320

The probes are trained on different Unet blocks and layers and have varying performances because of their difference in location, number of features, and output size (see Table 1). We choose the probe trained on Decoder 4 layer 1 because it produces the biggest outputs and achieves the highest performance metrics. Here are the performance metrics we consider:

Dice coefficient is a similarity measure that quantifies the similarity between two binary images. We use the Dice coefficient to compare the probe performances for salient object detection. The Dice coefficient is defined as:

$$\text{Dice} = \frac{2 \times \text{area of overlap}}{\text{total area}}$$

The Dice coefficient ranges from 0 to 1, where:

- 0 indicates no overlap between the binary images (complete dissimilarity).
- 1 indicates complete overlap between the binary images (complete similarity).

Spearman's rank correlation coefficient measures the strength and direction of association between the ranks of pixel intensities in two images. We use rank correlation to compare the probe performances for depth and shading prediction.

The value of Spearman's rank correlation coefficient (ρ) ranges from -1 to 1:

- ρ = 1 indicates a perfect positive monotonic relationship (as pixel intensities increase in one image, they also increase in the other image). We want our probing results show high rank correlation with the labels.

- ρ = -1 indicates a perfect negative monotonic relationship (as pixel intensities increase in one image, they decrease in the other image).

- ρ = 0 indicates no monotonic relationship between the pixel intensities of the two images.

Pearson's linear correlation coefficient measures the strength and direction of the linear relationship between two images. We use linear correlation in addition to rank correlation to compare the probe performances for depth and shading prediction. We want our probing results show high linear correlation with the labels.

Insert data table for comparing the probes' performance for the final report.

3 Results

(not required by the report checkpoint, but putting stuff down to frontload for the final report)

The probes are given the LDM's internal representation as inputs while the models are given intermediate images as inputs. To illustrate the LDM's internal encoding of 3D information, we compare the results from the probes with the image models (TRACER, MiDaS, and Intrinsic). Initially, due to the "noisy" nature of the intermediate images, these models struggle to generate accurate predictions. However, as the denoising steps progress, the models gradually improve in accuracy, coinciding with the point at which human observers start to discern relevant information more clearly.

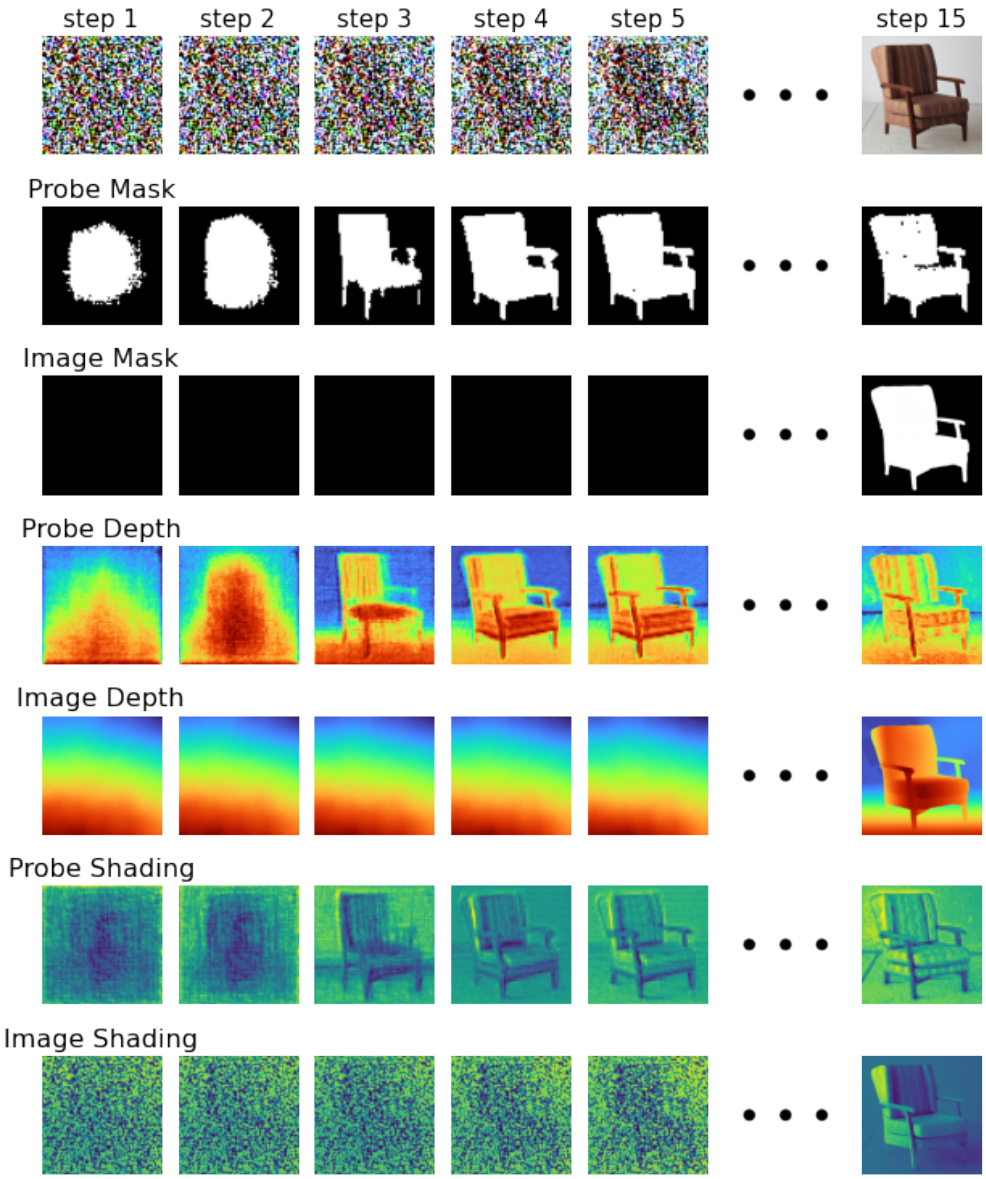

Figure 4: Probing Workflow

4 Discussion

5 Conclusion

6 Contributions

Everyone has worked on the probing experiments, and worked together to write the reports.

Karina explored existing texture models for potential probing of image texture. Karina investigated the Intrinsic model deficiencies. Karina edited the for loop in the .py file to execute a specific layer for a specific step.

Atharva created proof of concept code that we are able to classify images using a pre-trained model (on ImageNet), VGG-16. Atharva integrated VGG-16 wtih the code from Zelong to input intermediate images and output object category classification probabilities.

Ester modified existing code for depth probe to probe for shading using Intrinsic. generate full plot for a specific image. This plot compares all 15 steps of the generated image and their mask, mask_syn, depth, depth_syn, shading, shading_syn. Used Zelong's updated CUDA memory-efficient code to generate the intermediate steps for the image, mask, depth, and shading.

Zelong rewrote some parts of the code and solved the CUDA issue. The team used this CUDA memory-efficient code to generate the intermediate steps for the image. Zelong investigated the deficiencies in the Intrinsic model. Zelong designed the diagram for developing the probing classifier.

7 Appendix

7.1 Copy of Q2 Project Proposal

7.1.1 Broad Problem Statement

The ability of latent diffusion models (LDMs) to generate realistic images from textual descriptions has seen remarkable advancements. Even when trained purely on images without explicit depth information, they typically output coherent pictures of 3D scenes. These models have the astonishing capacity to create detailed, coherent scene representations.

However, their ability to represent depth and saliency within generated images remains unclear. Our project aims to delve into the diffusion process of LDMs, unraveling how they internally represent and process scene geometry.

This investigation is crucial as it not only enhances our understanding of AI's interpretive capabilities but also paves the way for further advancements in image synthesis. Existing research primarily focuses on the output capabilities of these models, leaving a gap in

comprehending their internal processing mechanics.

7.1.2 Narrow Problem Statement

Previous work mostly focused on using diffusion models to produce higher-quality results, particularly in image generation. In our Quarter 1 Project, we focused on simplifying the diffusion process, looking at the process of how DDPMs work on two-dimensional distributions, specifically how the Gaussian noise is removed in the reverse diffusion process.

In our Quarter 2 Project, we will similarly focus on the diffusion process, but we are now looking at their internal representations of scene geometry as noise is being removed. Even though diffusion models are trained on 2D images, we want to show that they encode some internal representations of depth and saliency even in early denoising steps before the human eye can detect anything other than random noise.

Previous work has attempted to answer this question and has found that there is depth information that emerges in early denoising. Baranchuk et al. (2021) extrapolated the intermediate activations of a pre-trained diffusion model for semantic segmentation. Their high segmentation performance reveals that the diffusion model encodes the rich semantic representations during training for generative tasks. Our work shows that the internal representation of LDM also captures the geometric properties of its synthesized images.

7.1.3 Primary Output Statement

Our primary output will be a research paper and a website to showcase our results. In our research paper, we will explain the motivation, methods, results, and discussion of our project. We should include images that demonstrate the LDM's ability to separate foreground from background in the early denoising stages. In contrast, image segmentation and salient-object detection models perform poorly on noisy images. We will compare the results from the LDM and the other methods to show that diffusion models can capture the internal 3D representation of a scene. On our website, we will display the most important information from our research paper with an emphasis on visualizations. To take advantage of the website format, we can include a tab for additional comparison examples and other interesting visualizations. As a stretch goal, we can implement an interactive widget that generates an image according to the object mask inputted by the user.

7.1.4 Data

For this project, our dataset consists of 617 images (512 pixels x 512 pixels) generated from Stable Diffusion v1.4. If we want to show the internal representation of our diffusion model through salient object detection, then our labels dataset consists of the salient object mask outputs from applying TRACER (https://github.com/Karel911/TRACER) to our generated images. If our goal is depth estimation, then our labels dataset consists of the monocular depth estimation outputs from applying MiDaS (https://github.com/isl-

org/MiDaS) to our generated images. We have successfully obtained all training and testing datasets, but we plan to generate more images using different prompts according to the directions we hope to explore.

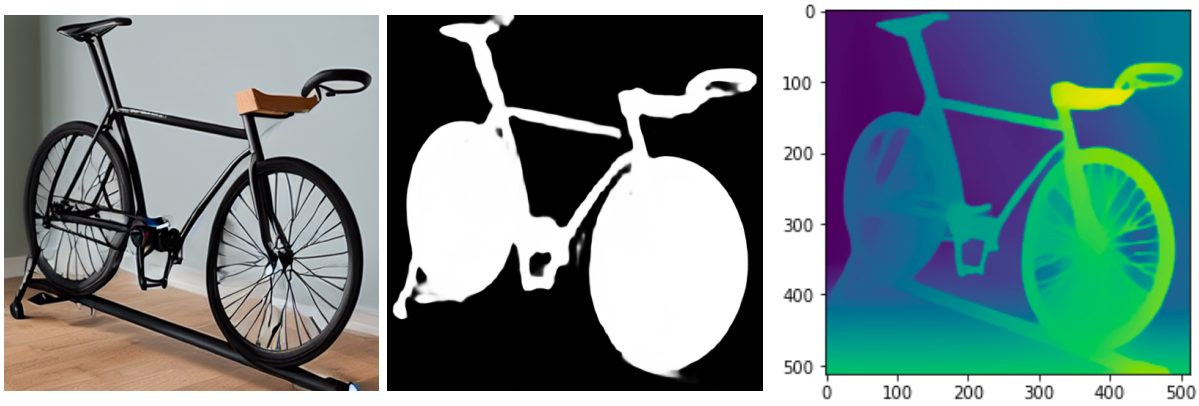

Figure 8: Three images of a bike: (a) *bike.png* - An image of a bike generated by Stable Diffusion v1.4 using the prompt "Lapierre Pulsium 600 FDJ Road Bike 2017" (b) Salient object detection mask output of *bike.png* using TRACER (c) Depth output of *bike.png* using MiDaS

References

Baranchuk, Dmitry, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. 2021. "Label-Efficient Semantic Segmentation with Diffusion Models." *CoRR* abs/2112.03126. [Link]

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

Chen, Yida, Fernanda Viégas, and Martin Wattenberg. 2023. "Beyond Surface Statistics: Scene Representations in a Latent Diffusion Model."

Lee, Min Seok, Wooseok Shin, and Sung Won Han. 2022. "TRACER: Extreme Attention Guided Salient Object Tracing Network."

Ranftl, René, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer."

Appendices

A.1 Training Details

A.2 Additional Figures

A.3 Additional Tables