Novel View Synthesis from Single Images with Tiny Latent Diffusion Models

Ethan Chun
Electrical Engineering and Computer Science
Massachusetts Institute of Technology
elchun@mit.edu

Abstract

Diffusion models show great potential to improve scene understanding in robotics applications. However, the computational cost of training and deploying state of the art models hinders their use in the real times settings found across robotics. We recognize that the fundamental goals of current diffusion models and robotics are misaligned; while the former focuses on image quality, robotic applications prioritize efficiency. To bridge this gap, we present a scaled down version of the popular Latent Diffusion Model, trained on the SRNCars dataset. Inspired by work on novel view synthesis with latent diffusion models, we add image and pose conditioning, allowing for the generation of novel views of unseen objects from a single input image. Our model reduces the number of trainable parameters from 860 million in state of the art models to less than 1.7 million parameters and trains in less than four hours on a consumer GPU. We demonstrate sweeps of novel views conditioned on a single input and offer commentary on how large diffusion models may be simplified for use in robotics applications [1]

1. Introduction and Motivation

Scene understanding is a longstanding problem in the field of robotics. In order to interact with the environment, any autonomous agent must understand exactly what the environment contains. Unfortunately, an agent often has a limited field of view full of occlusions, motion, and uncertainty, each making robust decision making extremely challenging. We postulate that novel view synthesis techniques from conditional latent diffusion models [3, 8] may hold the key to building more robust autonomous systems. These can allow an agent to infer the environment geometry, allowing the agent's decision processes to focus on higher level objectives.

While promising, the enormous compute capabilities [5] required to train these systems hinder their application to robotics. Furthermore, while conventional latent diffusion models prioritize image quality, robotics tasks have minimal quality requirements, instead focusing on computational efficiency. In this work, we remedy both challenges by building a tiny version of a classical conditional latent diffusion model. As shown in Fig. 1, our model takes a single conditioning image and a set of transforms as input, and produces novel views of the conditioning object at these new camera transforms. We train on the SRNCars [6] dataset to simulate a single object training pipeline that may be used in robotic pick and place tasks. Working with 32×32 input images and reducing the number of parameters from Stable Diffusion's [5] 860 million to less than 1.7 million, our proposed method trains in less than four hours on a consumer GPU while still synthesizing novel views from a single conditioning image.

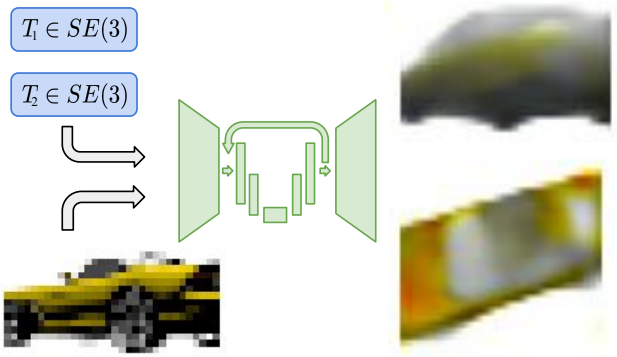

Figure 1. **Tiny Imaged Conditioned Diffusion Model –** Given a single conditioning image and a set of extrinsic camera transforms, our tiny latent diffusion model synthesizes novel object views from these new camera poses.

2. Related Works

Recent works in image diffusion models [2, 5] show the potential to generate or reconstruct images from conditional inputs like image or text prompts. Subsequent works lever-

[1] Code is available at https://colab.research.google.com/drive/1leJiaBDRsB_otVcb4GGfE0TlXTuxMJtD

age these models to generate novel views or 3D reconstructions [3] [8]. These offer promising generalization abilities, predicting potentially occluded features of a scene. This generalization makes conditioned 3D generative models particularly attractive for use in robotic manipulation.

2.1. Basics of Diffusion Models

As a basis for our method, we provide context on the development and components of a typical diffusion model. Diffusion models are a class of generative model that iteratively remove noise from a uniform normal distribution to generate a novel image. First introduced by Sohl-Dickstein in 2015 [7] and refined by Ho in 2020 [2] conventional diffusion models have shown promising results in a variety of image generation tasks. Subsequent work added conditioning tensors to diffusion models, allowing guided image generation. Unfortunately, these models were extremely resource intensive, owing to the the large image space which the denoiser must learn to denoise. In 2021, Rombach proposed a solution to this issue by running diffusion in the latent space of an autoencoder, rather than the much larger pixel space [5]. The enabled diffusion models to scale rapidly into the state of the art image generations models that we see today.

Consequently, diffusion models contain three main components: A variational autoencoder which transforms the image to and from the diffusion latent space; a denoising network which iteratively removes noise from the latent code; and a conditioning mechanism to allow image guidance. In this work, we design and implement a tiny version of each of these components in order to create lighter weight model.

3. Method

We implement a miniaturized version of Rombach et al's Latent Diffusion Models (LDM) [5] with image and pose conditioning inspired by Liu et al's Zero-1-to-3 [3]. To do so, we build a simplified version of the variational autoencoder, denoising UNet, and conditioning mechanism. We assume a generated and conditioning image size of 32×32, and a 4×4 homogeneous transform matrix to denote relative camera pose. We design the network to run on a consumer GPU and train in less than four hours. Our goal is to preserve the essential features of LDMs while removing complexity irrelevant to deployment in scene understanding tasks.

3.1. Tiny Variational Autoencoder

Mirroring the LDM implementation, we construct the encoder and decoder stages in three stages. Working towards the center of the autoencoder, the first module is the down or up sampling stage. This consists of a series of repeated ResNet blocks and down or up sampling blocks. This is followed by an attention layer sandwiched between two ResNet blocks. Finally, we normalize the result and pass it through a convolutional layer to arrive at the latent dimension. We find that a latent of size $1 \times 4 \times 4$ is sufficient to retain the richness of our dataset.

The encoder predicts two tensors of the latent size. One contains the means, μ, and the other contains the log variances, $\log(\sigma^2)$, of a diagonal Gaussian distribution. To decode, we sample $\mu + \epsilon\sigma$ from this distribution, assuming a multivariate normal noise parameter $\epsilon \sim \mathcal{N}(0, 1)$, and pass this into the decoder.

Simplifying the LDM loss function, we enforce a pixel-wise L1 loss and the KL divergence of our distribution from a multivariate standard normal distribution:

$$L_{VAE} = w_{L1}|x - \hat{x}| + w_{KL}KL(\mathcal{N}(\mu, \sigma), \mathcal{N}(0, 1)$$

Where $w_{L1} = 1.0$ and $w_{KL} = 5.0$. We find that for this small dataset, it is not necessary to include a perceptual loss term (such as the LPIPS loss [10]).

Using a latent dimension of $1 \times 4 \times 4$, the autoencoder contains 634941 parameters and trains in as little as 50,000 iterations or less than an hour on an Nvidia RTX4090 GPU.

3.2. Tiny Denoising UNet

Similar to the variational autoencoder, we implement the denoising UNet as a series of repeated ResNet and down or up sampling blocks. Since the encoder and decoder will not be split apart, we include skip connections between each ResNet block of the encoder and decoder. To facilitate conditioning, we add spatial transformer [9] blocks between each ResNet layer in the encoder and decoder.

When included in the latent diffusion model, the UNet takes an input of size $1 \times 4 \times 4$ and collapses this into a single 64×1 latent code. We train with L1 loss on the predicted noise at each diffusion time step.

In total, the UNet contains 1043089 parameters and trains in 20000 iterations. We find this takes around 20 minuets on an Nvidia RTX4090 GPU.

3.3. Tiny Conditioning Mechanism

In contrast to modern LDMs, we do not use the CLIP encoder and embeddings [4] to condition the network. Instead, since we exclusively work on novel view synthesis, we only condition on an input image and a 4×4 homogeneous transform matrix that denotes the camera location of the desired image in the frame of the conditioning image. We take the simple route and run the conditioning image through our encoder to get a $1 \times 4 \times 4$ latent code. We flatten both the encoded image and the homogeneous transform matrix and concatenate both to get a $1 \times 1 \times 32$ conditioning vector which we feed into the spatial transformer blocks.

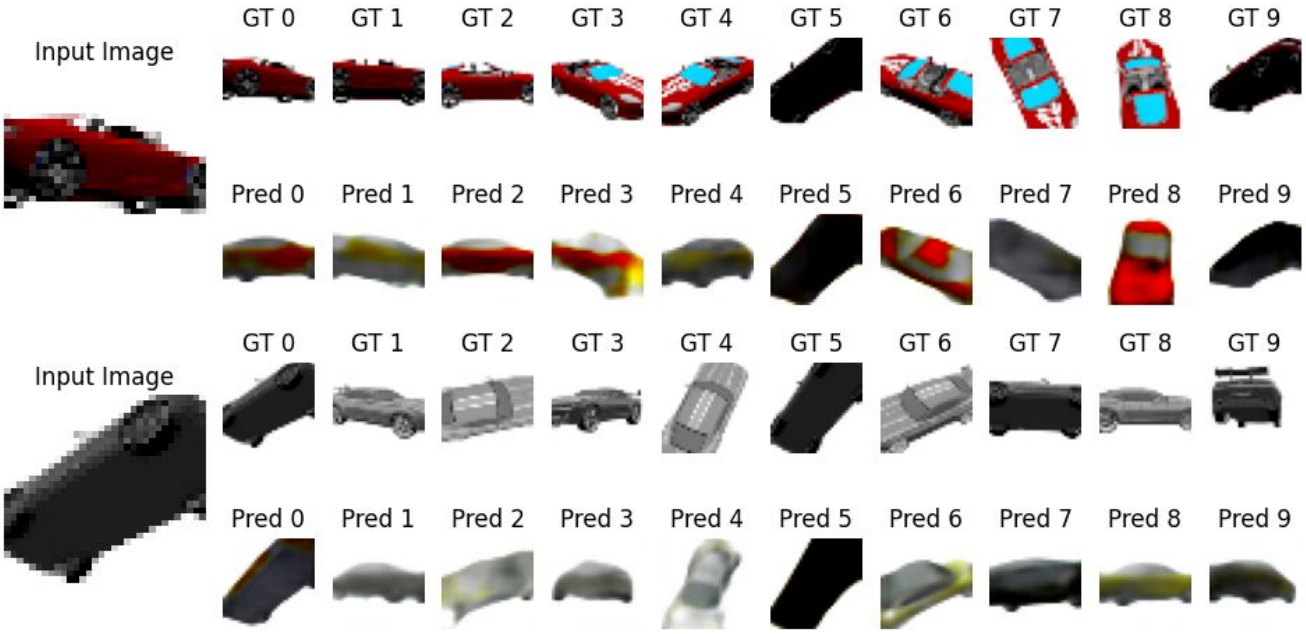

Figure 2. **Novel Views from Single Images –** Given a single input image (left) and a conditioning transform (not shown), our tiny latent diffusion model generates novel views (pred) of unseen geometry. Note that even when the model has no information on the top or back of the cars, it can still generate feasible predictions that resemble the unseen ground truth images.

4. Experiments and Analysis

We evaluate our model qualitatively by generating novel views of cars, conditioned on a single input image and the transform corresponding to that view. As shown in Fig. 2, our model is able to accurately generate a large number of the target images. Furthermore, we comment on characteristics discovered during the implementation and testing of the model. Given the brief structure of this write up, the commentary is not vetted and rigorously tested, but may prove useful to anyone building a tiny LDM.

4.1. Image Quality

In our tiny latent diffusion model, we find that the largest contributor to image quality is the quality of the latent space and reconstructions of the VAE. Given that every latent must pass through the decoder before being rendered, we think that the denoising UNet has comparatively little control over the quality of output images. Instead, in our model, the UNet almost exclusively worked as a control method to guide the latent to reflect the conditioning image and pose. The authors raise the question of whether full scale diffusion models function similarly and how replacing the autoencoder or UNet of a state of the art diffusion model may affect image generation quality and controllability.

4.2. Conditioning Mechanisms

In this implementation, we use the simplest conditioning mechanism and feed our image through our encoder while flattening the transpose matrix. As can be seen in Fig. 2, this produces reasonable results at this scale. Similar works encode the transform in spherical coordinates and concatenate this to the CLIP embeddings of the conditioning image [3]. However, for smaller scale models like ours, it appears that a simpler conditioning mechanism is sufficient.

4.3. Effect of Latent Size

We tested latents of sizes $4 \times 4 \times 4$, $2 \times 4 \times 4$, $1 \times 4 \times 4$, and $1 \times 8 \times 8$. Qualitatively, we found that the 4×4 latents required less regularization of the latent space and seemed to produce higher quality results. We also found that as we decreased the number of latent variables, the training times decreased while images quality remained constant. This indicates that a latent of size $1 \times 4 \times 4$ is likely sufficient to encode the SRNCars dataset. If one decided to train on portions of the ShapeNet dataset [1], for example, the latent size may need to be adjusted.

5. Limitations and Future Work

As may be apparent in Fig. 2 and Fig. 3, the model struggles to generate cars of different colors and, to a lesser extent, of largely different shapes. This may be due to an over representation of silver cars in the SRNCars dataset.

Future work may add a additional similarity metric to better enforce color consistency of cars. Additionally, it should be relatively simple to build a NeRF from the rendered views of the cars. Future work may extend this work to build a tiny single image to NeRF model which may be helpful in downstream applications.

6. Conclusion

We present a method to generate novel views from single images and target poses using a tiny latent diffusion model. We show promising results on the SRNCars dataset, generating a variety of unseen views using a single car conditioning image. Furthermore, we show that this model can be trained from scratch on a consumer gpu in less than 4 hours, making training tiny diffusion models more accessible. Given the low cost of training and the reduced complexity of this model, we hope that variations of this tiny conditional latent diffusion model find use in robotics, 3D reconstruction, and any other resource constrained task that may benefit from novel view synthesis.

References

[1] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 3

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 1, 2

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

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2

[6] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019. 1

[7] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[8] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Joshua B. Tenenbaum, Frédo Durand, William T. Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. In *arXiv*, 2023. 1, 2

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[10] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2

A. Additional Results

Figure 3. Additional cars generated at a variety of target poses.