Exploring Deep Generative Models to Conditionally Synthesize Street View Imagery

Praccho Muna-McQuay, Mindy Kim, Lisa Baek
Brown University

June 12, 2024

Abstract

In the domain of conditional image generation, diffusion models are the state-of-the-art. The applications of diffusion models are many, extending beyond images to speech and even video generation. Although there have been previous advancements with geo-transformations and implementations of Bicycle GANs for related tasks, there is a notable absence of Latent Diffusion Model (LDM) applications tailored for generating street view imagery from satellite data. Here, we present GeoLDM, an LDM that models the conditional distribution of Google Street View images given a set of corresponding geospatial features at that location. Further, we incorporate a pre-trained model released by SatlasAI for feature extraction and further conditioning of our model. To adapt the LDM to our geospatial task, we implement a variational autoencoder for ground image data and an interpolation head for satellite image embeddings to feed into a Denoising Diffusion Probabilistic Model (DDPM). Our method is capable of generating highly plausible street view images, and incorporates new features into the typical LDM framework as presented in previous works.

1 Introduction

With recent advances in artificial intelligence, developers have achieved significant improvements in modern image generation models. This study focuses on the text-to-image Stable Diffusion model, as introduced by Rombach et al. [6]. This model is capable of generating highly detailed images through diffusion in high-dimensional latent spaces. Similar to other generative models, diffusion models are also capable of conditioning on a set of inputs. Notably, the Stable Diffusion Model conditions on an encoded textual

description, which guides the diffusion process to produce outputs that are coherent with the input conditions.

Inspired by the success of the Stable Diffusion model and satellite-generative models, we seek to extend these methodologies to the generation of ground view images. We explore the efficacy of a diffusion model trained on geospatial data, investigating its ability to leverage satellite imagery and topographical data for generating realistic ground-level photographs.

Previous research has incorporated various conditioning methods and the implementation of a range of different generative and transformer models. Specifcally, the model proposed by the 'Geometry-Aware Satellite-to-Ground Image Synthesis' [4] utilizes Geo-Transformations, drawing on depth and semantic information from satellite-views to generate street view images using a Bicycle GAN framework. To enhance this approach, we propose two key innovations: first, we utilize a pre-trained foundation model generalized across numerous geospatial tasks, enabling richer feature representations for our conditioning process. This design choice is based on the rationale that a model pre-trained on a task closely aligned with ours will provide nuanced, better semantic guidance for GeoLDM, compared to general feature extraction methods previously employed.

Second, inspired by the work of Khanna et. al. in 'DiffusionSat' [3], we utilize Latent Diffusion Models (LDMs), which have demonstrated state-of-the-art performance in conditional image synthesis while significantly reducing computational demands relative to pixel-based diffusion models. Our goal is to demonstrate the effectiveness of employing an LDM in the generation of geo-imagery, conditioned on detailed representations of the desired image content.

2 Methodology

2.1 Data Preprocessing

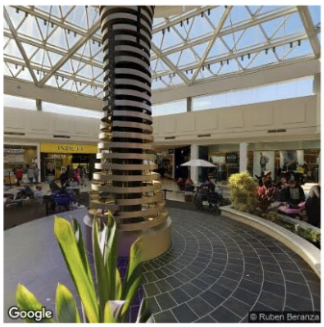
Without a pre-existing dataset for satellite and street view imagery, we utilized Google Maps and StreetView API to scrape images. We sampled longitude and latitude coordinates to collect corresponding pairs of satellite and street view images from each location. Due to the limited availability of other geospatial features during our scraping process, we restricted our conditioning variables exclusively to satellite imagery.

We limit our data to the United States, uniformly sampling points across the mainland and inputting these coordinates into the Google API. To ensure consistency and maintain focus on a specific area, we standardize the zoom level for all images and eliminate any Google watermarks. Additionally, we consider that the model should be invariant to rotations of feature maps, as they represent the same geographical region. To account for this, we incorporate random rotation into our satellite imagery:

Figure 1: same satellite images of Manhattan island, rotated

When initially training our autoencoder, we noted the limited diversity within our dataset, such as many of the images largely resembling one another and showing numerous photos of rural areas (i.e. lots of desert photos), which we attribute to this type of topography dominating much of the country's land coverage. To account for this, we collected more images, utilizing a normal distribution centered at cities with populations greater than 200,000. We note that this design choice was made halfway throguh the colelction of the data: thus, half of the dataset was sampled from the uniform distribution across mainland US, and the second half was sampled from the new urban distribution. To maintain this trend, we ensured that the testing dataset was also extracted in a similar manner.

After gathering images, we removed contaminated image pairs: any pairs where the scraped street view or satellite image did not load, coordinates point to locations on bodies of water, or any other inconsistencies that may interfere with the training process. However, we would like to note that there are circumstances that could not be accounted for, such as street view photos taken inside of buildings or hallucinating images such as the examples below.



In total, we collected 90,472 training image pairs, 1,674 validation image pairs, and 8,822 testing image pairs (where the distribution of uniform and city images are 50-50, reflecting the distribution in the training data).

2.2 Model Architecture

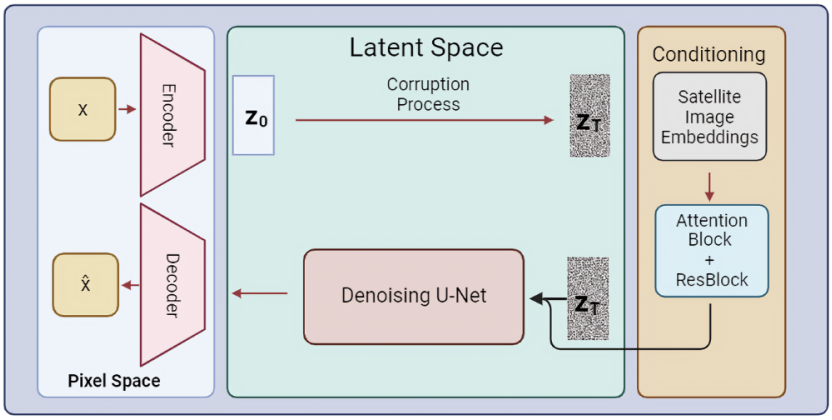


Figure 2: Above is a diagram of the GeoLDM architecture. The encoder, shown left, projects the image into the latent space, where corruption occurs, and a denoising U-Net is utilized to reverse the corruption process. Additionally, the conditioning factors, shown right - orange - adds additional factors for the U-Net to consider. The decoder, shown left, takes these outputs and projects the image back into regular pixel space.

Our GeoLDM architecture emulates a Latent Diffusion Model (LDM), similar to that of Stable Diffusion. We split up our training into multiple components, referring to the diagram above.

2.3 Ground Image Encoder & Decoder (blue, left)

Prior to training the diffusion part of GeoLDM(green, middle), we initially train an encoder and decoder on our Google StreetView images to generate a latent representation of the street view that the decoder can reconstruct the original image from.

This component consists of a convolutional variational autoencoder (VAE) trained on a combination of reconstruction, perceptual, and patch-based adversarial loss, in addition to a slight KL-loss. We use an L_1 distance for the reconstruction loss, which we complement with the perceptual loss that leverages LPIPS (Learned Perceptual Image Patch Similarity) – a distance metric based on perceptual similarity via relative activations within VGG [5]. At training step 50,000, we introduce a patch-based adversarial loss, which involves simultaneously training a patch-based discriminator that only penalizes structure at the scale of local image patches. This follows the methodology of the original stable diffusion paper [6]. The goal of this discriminator is to distinguish reconstructions

of street view images from original street view images. For a given street-view image x, our loss term for training the VAE (composed of an encoder $\mathcal{E}$ and decoder $\mathcal{D}$, with discriminator D_ω parameterized by ω) can thus be expressed as:

$$L_{\text{VAE}} = \min_{\mathcal{E},\mathcal{D}} \max_{\omega} L_{\text{rec}} + L_{\text{disc}} + \lambda D_{KL}(q_\phi(z|x) \parallel \mathcal{N}(\mathbf{0}; \mathbf{I}))$$

where

$$L_{\text{rec}} = L_1(\mathcal{D}(\mathcal{E}(x)), x) + L_{\text{LPIPS}}(\mathcal{D}(\mathcal{E}(x)), x)$$
$$L_{\text{disc}} = \log D_\omega(x) - \log D_\omega(\mathcal{D}(\mathcal{E}(x))$$

Together, these push the VAE to produce an encoding of our street-view imagery that captures minute details essential for accurate reconstruction, and provides a computationally efficient, low-dimensional latent space that GeoLDM can work in. By using a variational autoencoder specifically, we also ensure the data manifold is well formed. In line with the original stable diffusion paper, we set the weight λ on the KL term to $1e-6$.

During training, we pass these encodings through a corruption process, incorporating Gaussian noise into our images, before sending the noisy latent along with our satellite embedding into the denoising reverse process. Initially, we looked to implement a scaled-down version of the stable diffusion architecture, but experienced overfitting, attributed to a key difference in the original paper's versus our dataset. While the original architecture dealt with a vast variety of images and classes from the ImageNet dataset, our task deals with a very specific subclass and comparatively limited number of data. To resolve the overfitting, we shrank the bottleneck to shrink the embedding size - forcing the model to learn more meaningful representations that could still result in good quality reconstructions.

Once the encodings are passed through GeoLDM, the final outputs are then sent through our pretrained decoder, which reconstructs pixel images from the sample latent space, generating novel synthetic street view images.

2.4 Geospatial Feature Encoder (orange, right)

In order to produce the satellite embeddings for the conditioning step, we incorporate a pretrained model called Satlas AI [1], which was trained on a large-scale dataset for remote sensing image understanding to generate 'satellite embeddings' based on inputted geospatial data. Other previous works with satellite imagery rely on Satlas AI to extract meaningful representations from satellite imagery, rather than the image alone. We incorporate this model for a similar reason, as the predicted feature maps capture a richer variety of semantic information, assisting the conditional generation process within the DDPM.

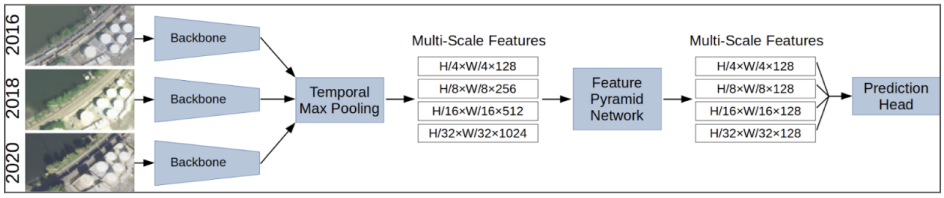


Figure 3: Architecture of the satellite embedding model, a pretrained model by Satlas AI hat generates satelilte embeddings based on inputted geospatial data.

The geospatial feature encoder consists of the backbone of SatlasAI (up to the Feature Pyramid Network part of the diagram), taking in single or multiple aerial images and outputting a multi-scale feature map. Additionally, we concatenated positional encodings of the latitudinal and longitudinal coordinates to provide further spatial information and conditioning to the model. This pertains to the idea that houses or other geographical features will be similar for coordinates that are closer to each other rather than farther away (e.g. New York buildings versus California buildings). The resulting feature embeddings were of dimensions 258×128.

There were a couple ablations we attempted during this stage of the model training. One change we compared was the difference between inputting the embeddings through an interpolation head, made up of a ResBlock followed by an AttentionBlock, versus sending the embeddings straight into the U-Net. We noted that the use of an interpolation head allowed the model to generate far more realistic imagery than without the head, which we thought was due to GeoLDM adapting the embeddings into a relevant, task-specific representation for the conditioning portion of our model.

We also decided to see if extracting more general image features would provide better, broader semantics of the satellite imagery. To explore this idea, we utilized a pretrained VGG model, a state-of-the-art deep convolutional neural network that provides a $4 \times 4 \times 512$ feature representation of the inputted image, which also gets sent in through the interpolation head to produce features that match the dimensions of the SatlasAI embeddings.

2.5 Diffusion Model (green)

GeoLDM generates high-quality images from a noisy latent, zt, of size $16 \times 16 \times 4$, conditioned on the encoded geospatial features (orange). The reverse denoising process was implemented with an augmented U-Net, as introduced by Ronneberger et. al in Biomedical Image Segmentation [7]. Composed of an encoder and decoder network, the U-Net utilizes convolutions and skip connections, which allow it to learn complex patterns with limited amounts of data. Within the layers of the U-Net, we chose to use a cross-attention mechanism, which allowed for increased flexibility in conditional image

generation. While implementing our model, we had originally intended on using the original implementation of U-Net, as Rombach et. al had done in their stable diffusion model. However, we ran into casting issues that were difficult to resolve. Additionally, we noticed that the calculations for diffusion would tend to 0 or NaN, which caused the model to produce black images. To fix this issue, we pivoted our approach to a re-implementation of the U-Net backbone architecture by Umar Jamil [2], which allowed us to successfully incorporate the U-Net as the backbone architecture for our DDPM. The goal of the U-Net is to predict the true noise ε_t added to the encoded latent z_0 at step t, given its current state z_t and condition y. During training, for a given batch of streetview and corresponding satellite images, we sample random time steps to add and remove noise from, optimizing the parameters of the U-Net to minimize the MSE between its prediction of the noise added at the corresponding step and the true noise. This gives us a loss term of

$$L_t = \mathbb{E}_{t\sim[1,T],\mathbf{z}_0,y,\epsilon_t}\left[\|\epsilon_t - \epsilon_\theta(F_t, t, y)\|^2\right]$$

Taking inspiration from the original stable diffusion model for the overall DDPM architecture, we configured the conditioning portion of the diffusion process, which allowed the model to condition on satellite image features rather than labels as the original Stable Diffusion model had. These changes allow for the model to train efficiently on the satellite embeddings. One additional ablation that we successfully implemented within our model was Classifier-Free Guidance (CFG), which acted as a 'conditioning dropout,' and removed our conditioning information – the satellite embeddings. The derivation of the formula is included below.

$$\begin{aligned}
\nabla_{\mathbf{x}_t}\log p(y|\mathbf{x}_t) &= \nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|y) - \nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t) \\
&= -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\Big(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\Big) \\
\bar{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t, y) &= \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - \sqrt{1-\bar{\alpha}_t}\, w\nabla_{\mathbf{x}_t}\log p(y|\mathbf{x}_t) \\
&= \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) + w\big(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\big) \\
&= (w+1)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - w\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)
\end{aligned}$$

Figure 4: formula for CFG, provided by Lilian Weng

Introducing CFG allowed the model to improve the quality of the produced pixel images by generating images unconditionally. The only modification necessary for this in the training process was creating a "null" condition for a given satellite image embedding and positional encoding, which involved setting their components to zero, which we justify the former by the mean of the feature maps being close to zero in our dataset. This null condition randomly replaces the true condition with probability 0.2, and allows us to learn the unconditional noise prediction. It is important to note that introducing CFG trades

diversity for fidelity, but based on qualitative analysis, many of the image reconstructions were not only of higher quality but also maintained relevant features from the satellite images. Particularly due to the size of the output images, we decided that it would be better to focus more on the quality of the images rather than just on the incorporated features.

3 Results

Our loss converged at around 30 epochs, with the MSE loss stabilizing at 0.229. This is an example of one of our results from training with the SatlasAI embeddings, showing the fed in satellite image (left, top), the ground truth API street view image (left, bottom), and the generated street view image from GeoLDM (right).

3.1 SatlasAI versus VGG Embeddings

To determine the efficacy of SatlasAI in producing embeddings for the inputted satellite images, we attempted the same task except fed VGG embeddings into the interpolation head instead:

We can see that there are certainly some similarities between the generated and ground truth images, with certain parts of the satellite images being represented within the corresponding generated street view image. However, the quality is not always clear, and there are quite a bit of differences that involve the object features. More results can be seen in the Appendix.

We concluded that qualitatively, SatlasAI's images seem to better capture topographical information given by the satellite images, which could be more meaningful results given the outlined task. However, we noted that VGG embeddings seem to outperform the SatlasAI embeddings in object generation, such as cars or houses.

Looking at the training loss over time, there does not seem to be a major difference between the model training losses using the SatlasAI embeddings versus VGG embeddings.

A couple of theories as to why SatlasAI qualitatively outperforms VGG include the fact that SatlasAI was trained on multiple geo-spatial tasks and mainly satellite images,

which means the embeddings may translate better than the generalized VGG features when attempting to produce street view images with the satellite information. Additionally, the VGG model may just be superior when it comes to object detection tasks, which is a crucial part of our model as the embeddings need to contain information about houses and other objects in our images, but does not do well when attempting to simply outline the geographical details in the generated image. However, these results are not conclusive, and more research needs to be done to actually prove these hypotheses to be true. Refer to Appendix A for additional results.

3.2 Classifer Free Guidance

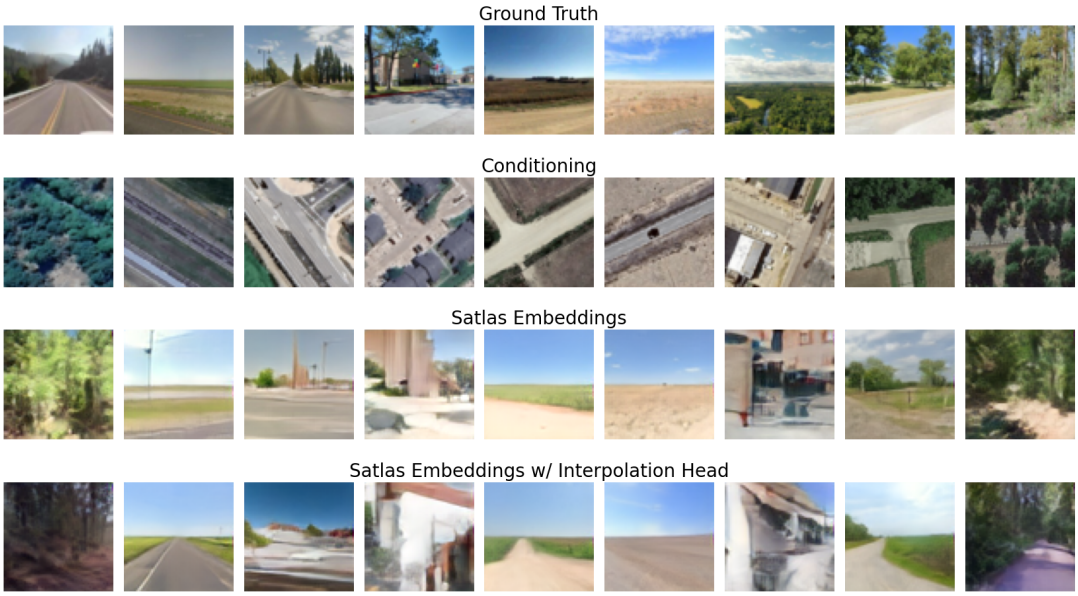
Additionally, we evaluated the importance of Classifier-Free Guidance (CFG) on the performance of GeoLDM (using SatlasAI Embeddings) by sampling with different CFG scales. Again, CFG is the 'conditioning dropout' that removes the satellite embeddings, or the conditioning information, at random times to further improve the quality of the outputted images. The following figure displays these results for nine satellite conditions.

Though it is difficult to tell any clear differences among these photos at different CFG scales, with no CFG, the sample quality (in terms of fidelity) is arguably worse across all these examples. As the CFG scale increases, we also see some evidence of greater adherence to the conditioning, as expected. For instance, in the first sample, we see at a

CFG scale of 5 and 7 that a body of water that is present in the conditioning and ground truth is visible in the generated image, whereas at a smaller CFG scale, that landmark is not evident in the generated image. The increase in fidelity is also observed in the second to last sample as the building appears more realistic at higher CFG scales.

3.3 Efficacy of the Interpolation Head

In our ablation studies, we also chose to explore the efficacy of the satellite interpolation head. To reiterate, the interpolation head is a trained residual and attention block, on top of the obtained satellite image embeddings. The reasoning behind this was to allow the model more plasticity in how it could use these embeddings from the pretrained, but frozen Satlas AI model, which we predicted would better serve the generation process. Without the interpolation head, these embeddings are passed directly to the UNet. Qualitative results are presented in the following figure.



From just looking at the generated images, it is again difficult to conclude a susbtantial advantage from including the interpolation head. However, in the second and eighth set of inputs, it could be argued that the model trained with an interpolation head produces more convincing results, including roads that appear in the corresponding satellite images.

3.4 Metrics

Initially, we sought to aim for more quantitative methods of evaluation for our model, utilizing metrics such as Frechet Inception Distance (FID) and Structural Similarity Index Measure (SSIM) to numerically determine the realism and diversity of images generated by GeoLDM. However, FID requires hundreds of thousands of images to compute on as fewer images result in much higher scores, and our goal was to generate plausible street

view images, meaning they did not have to look exactly like the original image, and, therefore, rendered SSIM to not be very representative of the model's performance. Thus, we chose to focus on more qualitative measures of our work, which one can determine for themselves using the examples above, as well as in Appendix A.

4 Future Directions

Within our implementation of CFG, we implemented static thresholding, following the formula included above. However, with recent advances in diffusion models, many papers focus on different forms of diffusion sampling: dynamic thresholding. Dynamic thresholding leverages high guidance weights and generates more photorealistic and detailed images than previously possible. With this implementation, we hope to decrease our FID score even further. The following paper below proposes more implementations that may improve our model's performance [8].

A Appendix

Trained with Satlas AI embeddings:

Fig: Satellite Images

Fig: API Street View Images

Fig: GeoLDM Generated Street View Images

Fig: Satellite Images

Fig: API Street View Images

Fig: GeoLDM Generated Street View Images

Trained with VGG embeddings

Fig: Satellite Images

Fig: API Street View Images

Fig: GeoLDM Generated Street View Images

Fig: Satellite Images

Fig: API Street View Images

Fig: GeoLDM Generated Street View Images

References

[1] Allen Institute for AI. (2023). satlaspretrain_models. GitHub. https://github.com/allenai/satlaspretrain_models

[2] Hyproj. (2023). Pytorch-stable-diffusion. GitHub. https://github.com/hkproj/pytorch-stable-diffusion?tab=MIT-1-ov-file

[3] Khanna, S., Liu, P., Zhou, L., Meng, C., Rombach, R., Burke, M., Lobell, D., & Ermon, S. (2023). *DiffusionSat: A Generative Foundation Model for Satellite Imagery.*

[4] Lu, Xiaohu & Li, Zuoyue & Cui, Zhaopeng & Oswald, Martin & Pollefeys, Marc & Qin, Rongjun. (2020). Geometry-Aware Satellite-to-Ground Image Synthesis for Urban Areas. 856-864. 10.1109/CVPR42600.2020.00094.

[5] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, & Oliver Wang. (2018). *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric.*

[6] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models.*

[7] Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation.*

[8] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding.*