Synthetic Generation of Satellite Images with Roads using ControlNet for Stable Diffusion

Ram Goel
MIT

Kerb: ramkgoel@mit.edu

Sumedh Shenoy
MIT

Kerb: sshenoy@mit.edu

Abstract

We present three ControlNet-based diffusion models—RoadNet, DilatedRoadNet, and ReducedRoadNet—that synthesize satellite images based on a user-given text input and road mask (a binary mask representing where roads are in an image), which we define as our prompt and input condition, respectively. We then qualitatively evaluate the performance each of these models based on prompt and condition fidelity, and compare the results. We find that ReducedRoadNet—which is obtained by training with a modified ControlNet training procedure that increases the prompt dropout rate during training—yields best results, matching the condition fidelity while improving the prompt fidelity of RoadNet and DilatedRoadNet.

1. Introduction

Considerable work has recently been done in the realm of employing machine learning techniques for analyzing synthetic satellite imagery data [1, 3, 5]. However, labelling large amounts of satellite imagery data with important features (e.g. roads) is manual and very resource-intensive. Consequently, an important problem that arises is the ability to generate such *synthetic* labelled satellite data in a scalable manner. This would provide the ability to amass substantial quantities of data that can be utilized for training satellite image machine learning models effectively.

In order to increase the size of a high-quality dataset for the training of supervised machine learning algorithms for satellite image analysis, it becomes imperative to generate satellite image data accompanied by their corresponding *road mask* (illustrated in Figure 1). The road mask serves as a visual representation exclusively highlighting the road structures within each generated image. By providing both the satellite image data and the associated road masks, we ensure the availability of crucial features necessary for training the algorithms to recognize and identify roads accurately.

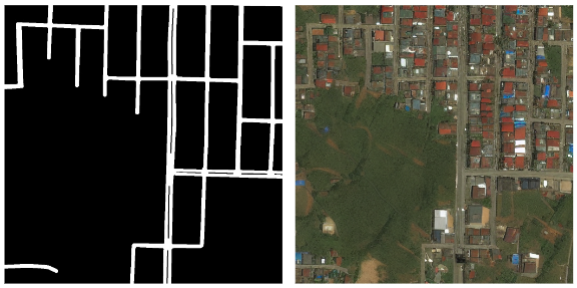

Figure 1. *Left: Example of road mask, Right: An image generated by our RoadNet model (Table 1) given this road mask, and a text prompt "rustic village".*

Our main idea is to use ControlNet [10], a neural network architecture that provides a backbone to an existing conditional diffusion model by providing the ability for further conditioning. ControlNet enables spatial conditioning for training diffusion models, thereby providing greater control over the generated images. This is key to our work, as the ControlNet architecture allows us to guide Stable Diffusion via road masks to synthesize satellite images. Then, the corresponding input road mask is simply the label of the generated satellite image, giving us already labelled data!

1.1. Research Problem and Goals

Our research goal is to generate satellite image data, drawing from two key inputs: (1) a road mask, and (2) an optional text prompt that offers additional specificity regarding the desired appearance of the generated image.

We will train ControlNet as a backbone to Stable Diffusion [7], using road masks as the additional input condition beyond just the text prompt. This road mask is what will guide the diffusion process spatially beyond text. We will train three such ControlNet models, with varying input datasets and training procedures. Training details are described in Section 3.1 and Section 3.2.

Figure 2. *Inpainting Stable Diffusion example. Left: inputted road mask, Middle: inputted full image, Right: outputted Stable Diffusion generated image.*

We will compare metrics across our models, including quality, prompt fidelity, condition fidelity, using test prompts and conditioning masks, and will analyze sources of variance across models, as detailed in Sections 3.3 and 4. Using this analysis, we will suggest methods to improve our models.

2. Related Work

Text-to-image generative modelling is a well-explored field, including many ideas not using diffusion. Stable Diffusion with Inpainting is a method for conditional image generation that can be used for the task of generating synthetic satellite images about roads. To use inpainting, the input is a text prompt and mask, along with a full input image. Given mask and image, uses Stable Diffusion to replace the parts of the image corresponding to the provided mask. Figure 2 provides an illustration of this technique applied to road masks. A key limitation of inpainting is that for inference, it requires a full image of input for each new generated image, not just a mask. This presents an issue for scalability, as it already requires large amounts of labelled data to generate new labelled data, and moreover, cannot synthesize new road shapes. In contrast, ControlNet-based methods only require as input a conditioning mask and a text prompt at inference time, which therefore does not require pre-labelled data at inference time, and can also synthesize satellite images with new road shapes.

Past methodology for text-to-image generative modelling that could be applied for this satellite imagery context includes GANs [6, 8, 9]. However, GANs face many convergence issues without extreme hyperparameter training, and also require very long training time, thus making them somewhat limited.

3. Methodology

Given a single image of a road mask—which is comprised of areas that are labeled as "road" being colored white, and all other regions being colored black, and an accompanying text description of the desired satellite image, we seek to generate a satellite image that is faithful to both the prompt and the road mapping. In particular, we wish to ensure that (a) no extra roads are added into the synthesized image that are not present in the provided road mask, and (b) no roads that are present in the road mask are *not* present in the synthesized image. Moreover, we seek to generate images that are able to display a wide variety of terrain and structures.

To that end, we train 3 different ControlNet models, each with a different dataset and/or training procedure. They all have the same road mask input condition. We later compare the performance of each model against one another; analyzing the strengths and weaknesses of the ControlNet architecture along both the axis of thickness of the road masks in the training data set and along the axis of prompt dropout rate, which we further describe in Sections 3.1 and 3.2.

3.1. Training Data Collection

To train our ControlNet models, we first prepare two datasets. The basis for both is the 6266 pairs of ground truth road masks and corresponding satellite images from the DeepGlobe 2018 Challenge [2]. We construct our first training dataset (henceforth referred to as *ControlRoads*) by first resizing both the ground truth road masks and corresponding satellite images from 1024×1024 to 512×512 pixels, and then using these resized images as our conditioning and target images, respectively. Then, for the captions associated with each condition and target image pair, we pass each satellite image through BLIP [4], an image-captioning model, and directly use the output as the caption for our mask-image-caption pairs.

Our second dataset is prepared similarily: however, rather than take the road masks as-is, we leverage the binary nature of the road masks to apply the OpenCV dilation function, using a 3×3 kernel of 1's, in order to increase the thickness of the roads. We generate captions identically to our first dataset. We will refer to this training data set as *ControlDilatedRoads*. The primary motivation behind such a dataset is the resulting roads in resized road masks being extremely thin, and therefore, slightly thickening these extremely thin areas provides a stronger control image that the model may potentially be able to learn more semantics from. We elaborate on experiments regarding this hypothesis in Section 4.

3.2. Training Procedures and Resulting Models

We provide two different training procedures: the first follows the training procedure outlined in [10], in which we pass pairs of conditioning images, target images, and captions (replacing a caption with the empty string "" with probability 50%) to our ControlNet architecture, and back-propogate the losses to update the weights. We use the standard ControlNet architecture described in [10], with Stable Diffusion 1.5 as our Diffusion model.

Model Name	Training Procedure	Train Dataset	Epochs
RoadNet	Standard ControlNet training procedure	ControlRoads	10
DilatedRoadNet	Standard ControlNet training procedure	ControlDilatedRoads	10
ReducedRoadNet	Modified ControlNet training procedure with a higher prompt dropout rate, as described in Section 3.2	ControlRoads	10

Table 1. *The three different ControlNets for Stable Diffusion we train, each with a different dataset and/or training procedure.*

For the second training procedure, we modify the original by increasing the rate of prompts randomly being replaced with the empty string to 70%, up from 50%. This is motivated by the homogeneity of the data; there are many common features and words that are shared in the descriptions, and as such, looking at fewer prompts could help reduce overfitting to features that are present in many of the prompts in the training dataset. We elaborate on experiments regarding this hypothesis in Section 4. With the the above datasets and training we procedures, we train three ControlNet models—*RoadNet*, *DilatedRoadNet*, and *ReducedRoadNet*—for Stable Diffusion, as outlined in the Table 1.

3.3. Evaluation and Metrics

We will evaluate our three models on the following three metrics:

1. **Image Quality**. We want to ensure generated satellite image data aligns with the provided road mask and exhibiting contextual coherence with the surrounding environment.

2. **Prompt Fidelity**. We will evaluate and compare how much our models are adhering to the prompt.

3. **Condition Fidelity**. We will evaluate and compare how much our models are adhering to the road mask condition.

To evaluate both the prompt fidelity and condition fidelity of our models, we take a qualitative approach: verifying by eye whether (a) the roads in the resulting image match up to the ones in the mask, (b) the resulting terrain/scene generated around the roads contains the all of the features expressed in the prompt, and (c) the resulting scene appears natural. We also qualitatively evaluate our models' robustness and ability to generalize past vocabulary that is densely represented in the training data set by applying these qualitative metrics to specific prompts, which we describe in Section 4.

4. Experimental Results and Discussion

As discussed in Section 3.3, we aim to evaluate our models' prompt fidelity and condition fidelity. We test these simultaneously, with four main types of test prompts and two main types of test conditions. These test prompts are as follows:

1. **No prompt**: We use the empty string "" as our prompt. This is important, as the quality of the resulting image helps gauge the robustness of the model for prompts, as described in [10]. For future use cases, a robust model is desireable: as primary use case for our ControlNet models is to be used to generate *reliable* synthetic data. Correspondig results can be seen in Figure 4.

2. **Terrain-type prompt**: We provide the model with a prompt that incorporates an "important" terrain feature we described above; i.e., "river", "mountain", "desert", etc. The purpose of this is to test that the models can synthesize these complex features contextually in the generated image. Corresponding results can be seen in Figures 3 and 6.

3. **Structural-type prompt**: We provide the model with a prompt that incorporates a man-made structural feature, such as city buildings, or farms. Corresponding results can be seen in Figures 3 and 6.

4. **Weather-type prompt**: we provide the model with a prompt that incorporates weather conditions. For now, we only test cloud incorporation, as other weather conditions are not conducive to satellite images. Corresponding results can be seen in Figure 5.

The test conditions are:

1. **Dense road masks**: We give the model a condition with a road mask we qualitatively define as "dense"—namely, where there are many roads spanning the image, which impose a significant constraint on the image. Corresponding results can be seen in Figure 3.

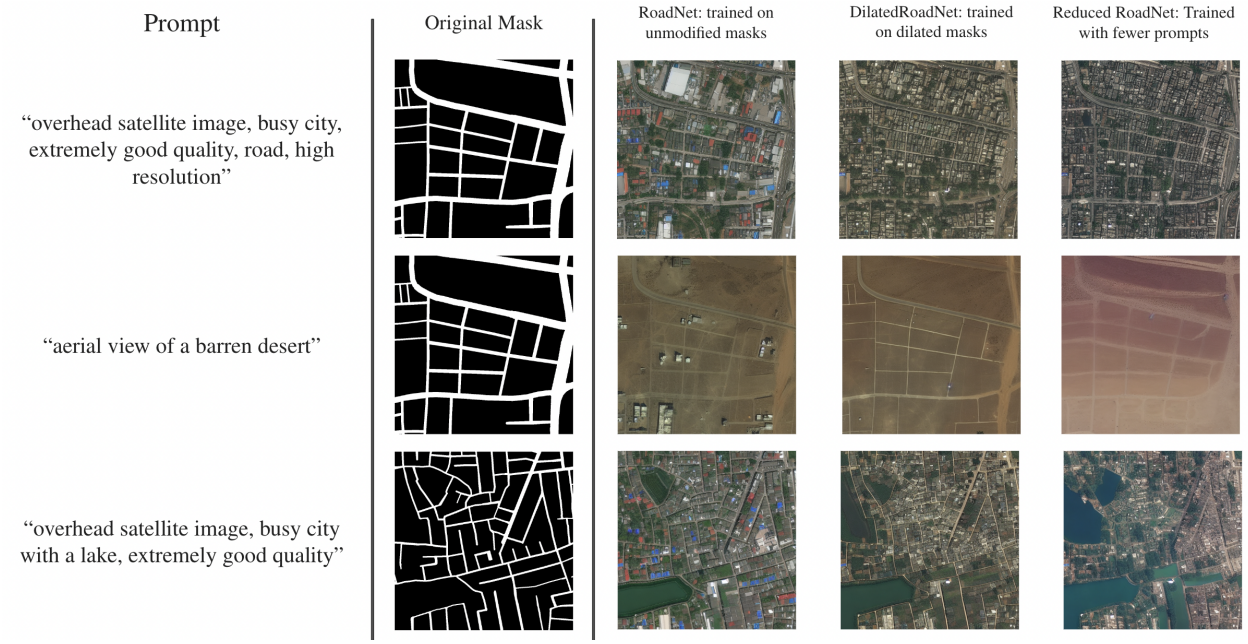

Figure 3. *Results from our three models RoadNet, DilatedRoadNet, and ReducedRoadNet on* ***dense*** *road maps. Observe that all 3 models achieve good performance on the dense maps, rendering natural looking images with high prompt and condition fidelity. Left column: Input text prompt. Middle column: Inputted original mask. Right column: Output for each of the corresponding input prompt and original mask.*

2. **Sparse road masks**: Similarly, we give the model a condition with a road mask that we qualitatively define as "sparse"—namely, there are not many roads spanning the image, meaning there is a significant portion of the image that is not controlled by the condition. Corresponding results can be seen in Figure 6.

Note that for test prompts, being terrain, structural, or weather-type is not mutually exclusive; rather, it is the opposite. We frequently test prompts that are multiple types to test how well our model is able to incorporate multiple such features into a single image; as being able to do so is critical for generating more complex images.

From Figures 3 and 6, we see that all three of our models produce *high-quality* images for both sparse and dense road masks; being able to generate natural-looking images while still preserving all of the roads (and not adding in new ones). Moreover, they appear natural, and perform well on terrain and structure prompt tests, reflecting both in the resulting images, as desired. Being able to generate said images from *only* a road mask is a substantial improvement from the previous Stable Diffusion with Inpainting approach.

However, we see clear issues in RoadNet and DilatedRoadNet when no prompt is given—as shown in Figure 4, they both produce highly unnatural results. This reveals that they are both overdependent on the given prompt, as they are unable to generate natural-looking images without sufficient guidance. On the other hand, ReducedRoadNet is able to produce natural looking images. This is likely due to the higher prompt dropout rate, as this helps the trainable encoders in the ControlNet model recognize semantics from the condition input mask, which ReducedRoadNet is able to do, whereas RoadNet and DilatedRoadNet cannot.

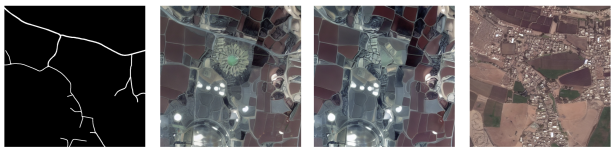

Figure 4. *Given an empty prompt, from left to right, we have: the input road map condition, and the output from RoadNet, DilatedRoadNet, and ReducedRoadNet. Note that RoadNet and DilatedRoadNet fail to generate a natural looking image.*

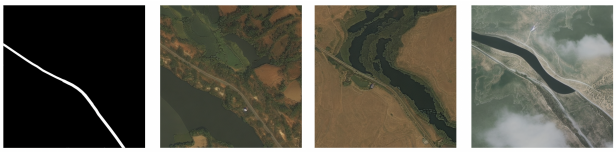

Figure 5. *Given the prompt "overhead aerial photograph, cloudy day, landscape with a river running through it", from left to right, we have: the input road map condition, and the output from RoadNet, DilatedRoadNet, and ReducedRoadNet. Note that RoadNet and DilatedRoadNet fail to include clouds.*

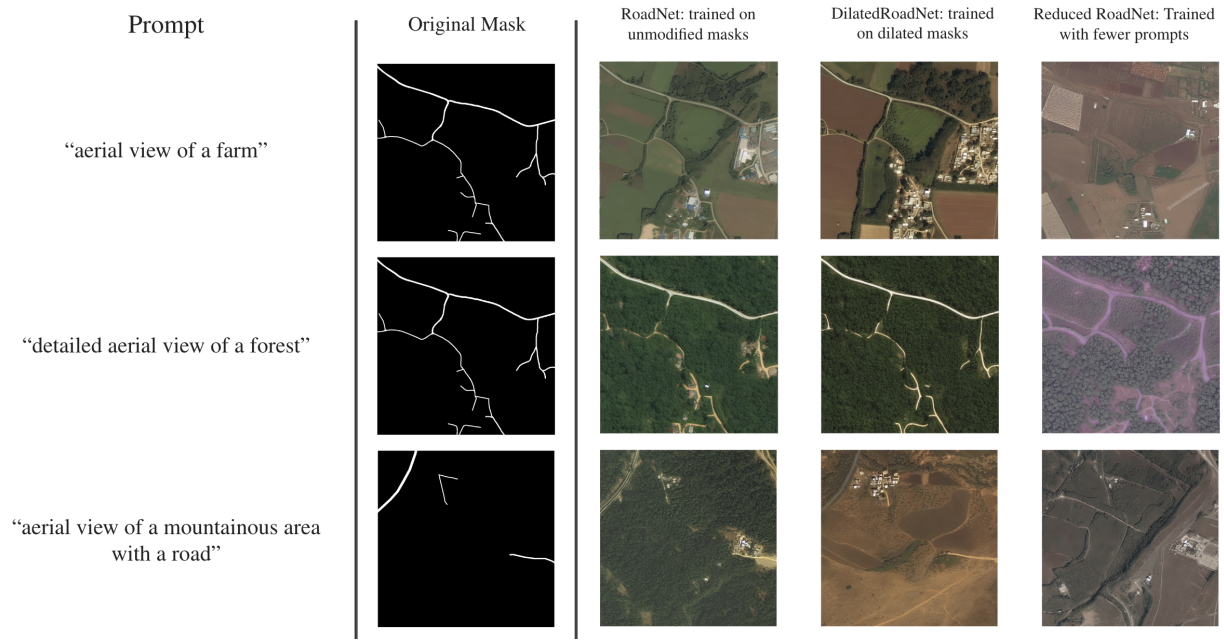

Figure 6. *Results from our three models RoadNet, DilatedRoadNet, and ReducedRoadNet on sparse road maps. Observe that all 3 models achieve good performance on the sparse maps, rendering natural looking images with high prompt and condition fidelity. Left column: Input text prompt. Middle column: Inputted original mask). Right column: Output for each of the corresponding input prompt and original mask.*

Thus, we find that the higher dropout rate contributes the robustness of our model, likely due to the small training set.

This difference in capability is further demonstrated when we proceed with weather prompt tests, as seen in Figure 5. When provided with the description of “cloudy” or “clouds”, both RoadNet and DilatedRoadNet completely fail to capture said feature. On the other hand, ReducedRoadNet is able to succesfully incorporate clouds into both images. Since the word “cloud” is minimally represented in the training prompt dataset, it is likely that RoadNet and DilatedRoadNet are overfitting to the vocabulary within the training dataset—evidenced by their ability to provide quality images of the terrain and structural prompts, as those features are well-represented in the train dataset. On the other hand, by increasing the dropout rate for prompts, ReducedRoadNet is able to avoid overfitting to the prompt train dataset, yielding better results.

Finally, across all experiments, we see negligible differences in quality between RoadNet and DilatedRoadNet, indicating that the dilation pre-processing does not have a significant impact.

5. Conclusion

We trained three different ControlNet models to generate high-quality satellite imagery, given an input condition of a road mask along with an optional descriptive text prompt. We desgined separate training procedure and training datasets for these three models, in order to test how much condition fidelity and prompt fidelity are affected by the probability of making a prompt empty and by thickening the roads in the road mask training dataset.

All our models produce generally high-quality images—the generated images are largely faithful to the road mask, and the models were able to synthesize complex surroundings. We found that ReducedRoadNet performed the best with respect to both prompt and condition fidelity, since we observed that RoadNet and DilatedRoadNet overfit to the training prompt data, lacking robustness, and not generalizing well to vocabulary they had not seen before. We also found that DilatedRoadNet and RoadNet produced very similar results, and thus dilation preprocessing does not provide significant improvement.

Future work could include testing different datasets and training procedures for the road mask conditioned ControlNet. Regarding the dataset, we could try changing the image captioning model, in order to potentially diversify the vocabulary seen, thus potentially affecting prompt fidelity. We found here that an increased dropout rate reduced overfitting to the prompt, so regarding the model, an intriguing question would be finding the optimal caption dropout rate. This is an open problem, and could have implications in the larger context of text-to-image generative models.

6. Individual Contributions

1. Sumedh Shenoy: trained the ControlNet models, and proposed the increase in prompt dropout rate. Both wrote and edited the reports and slides.

2. Ram Goel: preprocessed the datasets. Both wrote and edited the reports and slides.

References

[1] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. DeepGlobe 2018: A challenge to parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, jun 2018. 1

[2] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 2

[3] Rolf E., Proctor J., and Carleton T. et al. A generalizable and accessible approach to machine learning with global satellite imagery, 2021. 1

[4] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2

[5] Sharada Prasanna Mohanty, Jakub Czakon, Kamil A. Kaczmarek, Andrzej Pyskir, Piotr Tarasiewicz, Saket Kunwar, Janick Rohrbach, Dave Luo, Manjunath Prasad, Sascha Fleer, Jan Philip Göpfert, Akshat Tandon, Guillaume Mollard, Nikhil Rayaprolu, Marcel Salathe, and Malte Schilling. Deep learning for understanding satellite imagery: An experimental survey. *Frontiers in Artificial Intelligence*, 3, 2020. 1

[6] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis, 2016. 2

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1

[8] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2

[9] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, 2017. 2

[10] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 2, 3