GeoSynth: Contextually-Aware High-Resolution Satellite Image Synthesis

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Nathan Jacobs
Washington Univerity in St. Louis
{s.sastry, k.subash, a.dhakal, jacobsn}@wustl.edu

Figure 1. Satellite images synthesized by GeoSynth using OpenStreetMap for layout control and textual prompts for style control.

Abstract

We present GeoSynth, a model for synthesizing satellite images with global style and image-driven layout control. The global style control is via textual prompts or geographic location. These enable the specification of scene semantics or regional appearance respectively, and can be used together. We train our model on a large dataset of paired satellite imagery, with automatically generated captions, and OpenStreetMap data. We evaluate various combinations of control inputs, including different types of layout controls. Results demonstrate that our model can generate diverse, high-quality images and exhibits excellent zero-shot generalization. The code and model checkpoints are available at https://github.com/mvrl/GeoSynth.

1. Introduction

Imagine a scenario where you describe a scene and a layout and a realistic satellite image blooms into existence. Such kind of applications could assist in various remote sensing pipelines like urban planning, data augmentation, pseudo label generation for weakly supervised learning, etc. However, a single satellite image usually binds several spatial concepts into a unique image depicting complex and meaningful layouts. A scene on the ground, for example, may contain buildings, roads, intersections, crosswalks, trees, etc all placed together in a specific arrangement. This makes the problem of synthesizing realistic satellite images very challenging.

Recently, text-to-image models have rapidly redefined the realms of creativity and expression. When trained on vast datasets, they become capable of generating everything from photorealistic landscapes to fantastical creatures. In this regard, diffusion models [18, 35, 40] have shown impressive performance in a variety of tasks such as image generation [32, 38], image editing [2, 16, 21], video generation [3, 14], etc.

Similarly, the field of remote sensing has witnessed remarkable progress in various aspects, including imaging technology, accessibility of high-resolution data, and global-scale applications. Thanks to the development of large-scale foundational models [8, 20, 33, 42, 55], many remote sensing challenges have been tackled recently. The desirable properties of such foundational models have enabled domain-specific solutions in fields such as language [39, 53], sound [22], natural images [9, 45], etc. However, the majority of these machine learning-based methods fall short of utilizing the full potential of the satellite image modality [34]. Along the same direction, less attention has been given to satellite image synthesis in remote sensing. Existing approaches to this problem are either application-specific [15] or lack personalization capabilities [23].

A fundamental difficulty in using the already existing diffusion models for synthesis is that diverse and high-resolution satellite images are unseen during their large-scale training. Furthermore, despite their general abilities, they fail to synthesize multiple specific concepts in an image [26]. Recent lines of work addressing this challenge use a variety of techniques such as fine-tuning textual prompts [12], end-to-end training [36], manipulating the generation process [1], etc. ControlNet [54] has emerged as a promising architecture that learns to utilize information from an existing large-scale diffusion model while allowing it to condition it with a variety of controls. In this study, we use ControlNet to fine-tune Stable Diffusion (SD) [35] to synthesize satellite images.

As shown in Figure 1, this work aims to be able to synthesize realistic-looking satellite images, whose layout could be controlled via a reference image (for example OpenStreetMap (OSM) images), while style could be controlled using textual prompts. We additionally condition our models on geographic location using features extracted from SatCLIP [25], a model trained contrastively on satellite images and geographic location. By doing so, our model exhibits synthesis capability conditioned on the geography of a region. In addition to the OSM control, we test two other conditioning controls: Canny edge and Segment Anything (SAM) [24] mask, which can be directly obtained from raw satellite images. In the end, we have a suite of models namely, GeoSynth, which is capable of synthesizing satellite images that are optionally conditioned on layout, textual prompt and/or geographic location. The contributions of this work are threefold:

1. We use features extracted from ControlNet and SatCLIP for high-resolution satellite image synthesis.
2. We test the performance of three conditioning controls for synthesis: OSM image, Canny edge image, and Segment Anything mask.
3. We demonstrate excellent zero-shot capabilities of our models.

2. Related Works

2.1. Diffusion Models

Sohl-Dickstein *et al.* [40] first proposed physics-inspired generative models called diffusion probabilistic models to learn data distribution through parameterized Markov chains. Inspired by this, Ho *et al.* [18] demonstrated that diffusion models can generate high-quality images. In the following years, Stable Diffusion [35] was introduced which proposed training in the latent space of pre-trained autoencoders leading to low computation cost and high-quality conditional image synthesis. Another competitive model, Imagen [38] utilizes a powerful large language model (LLM) to yield photorealistic images conditioned on text while training the diffusion model in pixel space. Motivated by the impressive results of these works, there has been an explosion of numerous diffusion models for controlled image synthesis [10, 32] leading to diverse capabilities such as image editing [2, 16, 21, 49], image stylization [13, 19, 41, 43], and video generation [3, 11, 14].

2.2. Customization of Diffusion Models

Leveraging the power of existing diffusion models pre-trained on large-scale datasets, two lines of work have emerged. The first focuses on developing methods to personalize the pre-trained diffusion model to encapsulate custom concepts and domains [7, 26, 27, 29, 30, 36, 37]. For instance, Dreambooth [36] proposes personalization by learning from a small set of subject-specific images while preserving the prior of the pre-trained diffusion model. Other notable works propose techniques such as hypernetwork learning [37], modifying the cross-attention layers [26], apprenticeship learning from a large number of concept-specific experts [7], and utilizing the knowledge of existing multimodal representation space [27, 29]. The second line of work focuses on developing training-free or efficient methods to incorporate different types of conditions into powerful diffusion models [5, 48, 51, 54, 56]. ControlNet [54] proposes zero-convolution-based modules to incorporate additional conditions such as Canny edge [4], sketch, pose, etc. Uni-ControlNet [56] proposes to train separate adapters for two sets of controls: local controls (e.g., segmentation masks) and global controls (e.g., CLIP embeddings). Apart from these, a few recent works have focused on spatial layout-controlled image synthesis [5, 48, 51]. Inspired by these works, we utilize the spatial layout present in OSM images as one of our conditions while employing the flexible approach of ControlNet for our conditional image synthesis.

2.3. Satellite Image Synthesis

The literature on conditional satellite image synthesis is limited. Most of the previous works are focused on task-specific satellite image synthesis. For example, a recent work, EDiffSR [47] proposes to utilize diffusion models for the task of super-resolution of remote sensing images. Chen *et al.* [6] demonstrate the utility of features from diffusion model trained on hyperspectral remote sensing imagery, for the task of pixel-wise semantic segmentation. For the task of text-to-satellite image synthesis, [50] proposes a two-stage framework. First, a VQVAE-like framework [44] is trained to learn a codebook of visual representations for satellite imagery. Second, text-conditioned prototypes are learned to be utilized by VQVAE decoder to synthesize an image from text. A parallel work, DiffusionSat [23] learns satellite image generation conditioned on freely available metadata and sparsely available textual description. They demonstrate the impressive performance of their model on various downstream tasks such as super-resolution and in-painting. Different from their work, we use detailed textual descriptions along with spatial layouts for satellite image synthesis. We propose to incorporate all the conditioning modalities through ControlNet, hence preserving the existing knowledge base of stable diffusion. Additionally, we use a foundational location encoder model: SatCLIP, to incorporate geographic location for the synthesis process.

3. Background

Diffusion Models. Diffusion Models [18, 40] are a class of probabilistic models that learn to sample from a data distribution ($\mathcal{D}$) given numerous samples from that distribution. This is done by learning to denoise a variable sampled from a known prior noise distribution in a markov chain process. A popular choice for this noise distribution is the standard normal distribution. During training, given a noisy image x_t at timestep t, the objective of the diffusion model (ϵ_θ) is to predict the noise added at timestep $t-1$ to obtain that image. The criterion is given by:

$$\mathbb{E}_{x,c,t,\epsilon}\left[||\epsilon - \epsilon_\theta(x_t, c, t)||_2^2\right] \quad (1)$$

where $x \sim \mathcal{D}$, c is a conditioning modality, $t \in \{T, T-1, \ldots, \delta\}$ and $\epsilon \sim \mathcal{N}(0, I)$. c can be text, raw image, segmentation map, etc.

Latent Diffusion Models (LDMs). Since the dimension of the original data distribution in the case of images is high, the diffusion process is computationally expensive. LDMs [35] proposed to first encode the original samples into a low dimensional latent space and then perform the diffusion on the latent representations of the original samples. After a series of denoising steps, a decoder is used to reconstruct the original image from its latent representation.

ControlNet. ControlNet [54] is used to add additional conditioning controls to an existing neural network without having to fine-tune the original network. This is done by transforming the feature maps extracted from the existing neural network, into a feature that is conditioned on a given control. Each block of the ControlNet is connected to a zero-initialized layer, which ensures no noise is added during training.

4. Dataset

We built a dataset consisting of paired high-resolution satellite imagery and OSM images. We use a static representation of OSM in the form of images. The pairs were sampled randomly near ten major US cities, as shown in Appendix B. To improve coverage and reduce spatial bias, each sampling location is spaced at least 1 kilometer apart from one another. All the images downloaded are of size 512x512 pixels at an approximate ground sampling distance of 0.6m. We downloaded 90,305 image pairs and filtered pairs consisting entirely of bare Earth, water, or forest. After filtering, the dataset contained 44,848 pairs.

Figure 2. Each dataset sample consists of a satellite image, an OSM image, and an automatically generated textual description. Additionally, the dataset includes SAM masks for each satellite image.

We extended the dataset by captioning each satellite image using LLaVA [28], a recently released multimodal large language model (Figure 2). The prompt used for captioning was: "Describe the contents of the image". The captioning pipeline took 40 GPU hours to run on 2 NVIDIA A6000 GPUs. Lastly, we extracted the Canny edge image and the Segment Anything mask corresponding to each satellite image.

5. Method

Our goal is to train a suite of models that are capable of synthesizing satellite images (x) given a text prompt (τ), geographic location (l), and a control image (c). This is done by training diffusion models to learn the conditional distribution $p(x|\tau, l, c)$. To this end, we use Latent Diffusion Models (LDM), which have shown state-of-the-art performance in conditional image synthesis. Below, we describe the model architecture used and the implementation details of the training pipeline.

5.1. Architecture

We utilize a pre-trained LDM that comprises four primary architectural components. Firstly, an encoder that transforms raw images into a low-dimensional latent space. Secondly, a pre-trained CLIP text encoder [31] processes the raw text prompts and generates latent text vectors. Thirdly, the diffusion model has a U-Net based architecture con-

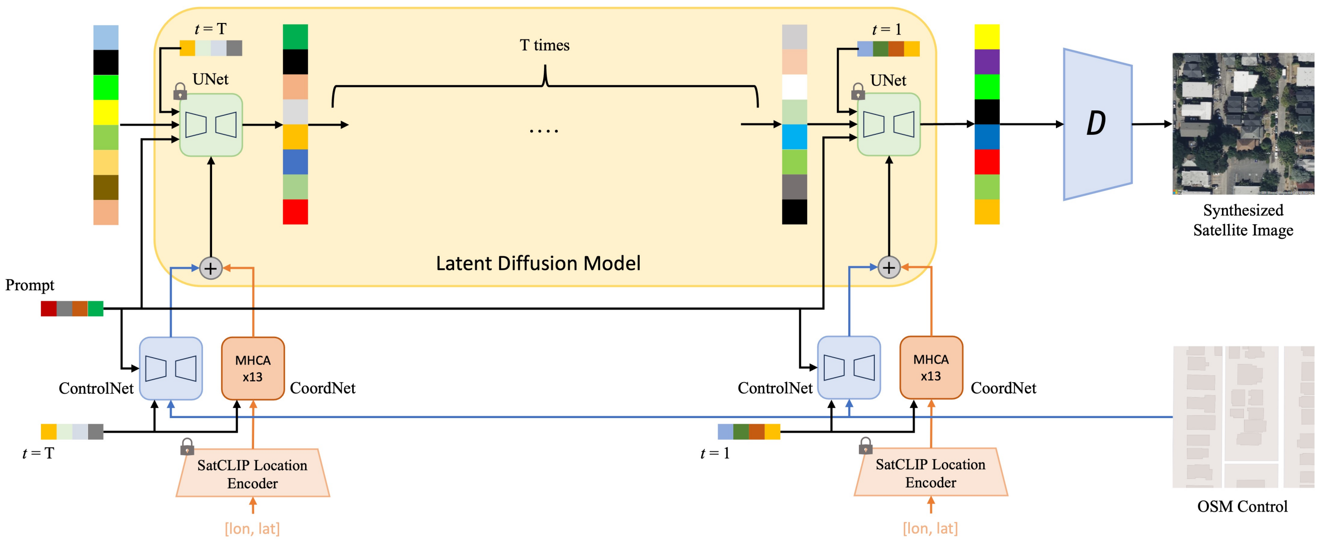

Figure 3. A high-level architecture overview of GeoSynth, which consists of a pre-trained LDM, ControlNet and CoordNet.

sisting of cross-attention blocks. Finally, a decoder reconstructs images given their corresponding latent vectors. The encoder and decoder of th LDM have a Variational Autoencoder (VAE) style architecture. During training, the diffusion process is used in the latent representation space of the raw images. The diffusion model learns to denoise a noisy latent vector at a given timestep conditioned on the text prompt.

As shown in Figure 3, we use ControlNet to incorporate a control image and fine-tune the pre-trained LDM. ControlNet is a zero-initialized neural network attached on top of an LDM, which transforms the feature maps of the LDM at each stage. The ControlNet architecture consists of 13 residual cross-attention blocks which take as input the control image, text prompt, and the diffusion timestep.

To incorporate geographic location as a condition, we first use SatCLIP [25] to extract location-based features. SatCLIP is a spherical harmonics-based location encoder that provides general-purpose location embeddings. It is trained using a contrastive learning framework with a CLIP-style satellite image encoder. We design a ControlNet-style cross-attention-based transformer, namely CoordNet, which processes the location embeddings. CoordNet consists of 13 layer multi-head cross-attention blocks which take as input the SatCLIP location-based embeddings and the diffusion timestep. Each cross-attention block in the CoordNet consists of a zero-initialized feed-forward layer. The features extracted from each of its blocks are added to the features extracted from the corresponding blocks of the ControlNet. These features are then added to the corresponding residual blocks of the LDM. During training, all the LDM components and the SatCLIP location encoder are frozen, as shown in Figure 3.

During inference, a noisy latent vector is sampled from a standard normal distribution. The diffusion model is then used to progressively denoise the latent vector over a series of T timesteps. The inputs from the CLIP text encoder, the ControlNet, and the CoordNet are used to guide the denoising process at each timestep.

5.2. Implementation Details

We use Stable Diffusion (SD) v2.1 as the pre-trained LDM. We use the same base architecture of ControlNet as used by authors of [54]. CoordNet consists of stacked 13 cross-attention blocks with an inner dimension of 256 and 4 heads. To improve training speeds, we precomputed the location-based embeddings of SatCLIP for each of the images and saved them on disk.

In total, we trained 14 variants of the model, including and excluding the conditioning modalities. Each variant of the model is trained on 2 NVIDIA RTX 4090 for a total of 100 GPU hours. We train the models using the Distributed-DataParellel routine in PyTorch. We use the Adam optimizer with a learning rate of $1e^{-5}$, gradient accumulation over 16 batches, and a batch size of 4 on each GPU. Following [54], we randomly mask out the textual prompts with a probability of 0.5 during training. This ensures that the model learns the semantic information present in the control images independent of the textual prompts.

We use three distinct metrics to evaluate the performance of the models. The first metric is the Frechet Inception Distance (FID) [17], which indicates the distance between the synthesized and ground-truth data distribution at the feature level. The second metric is SSIM [46], which measures the similarity between synthesized images and ground-truth samples at the pixel level. Lastly, we use the

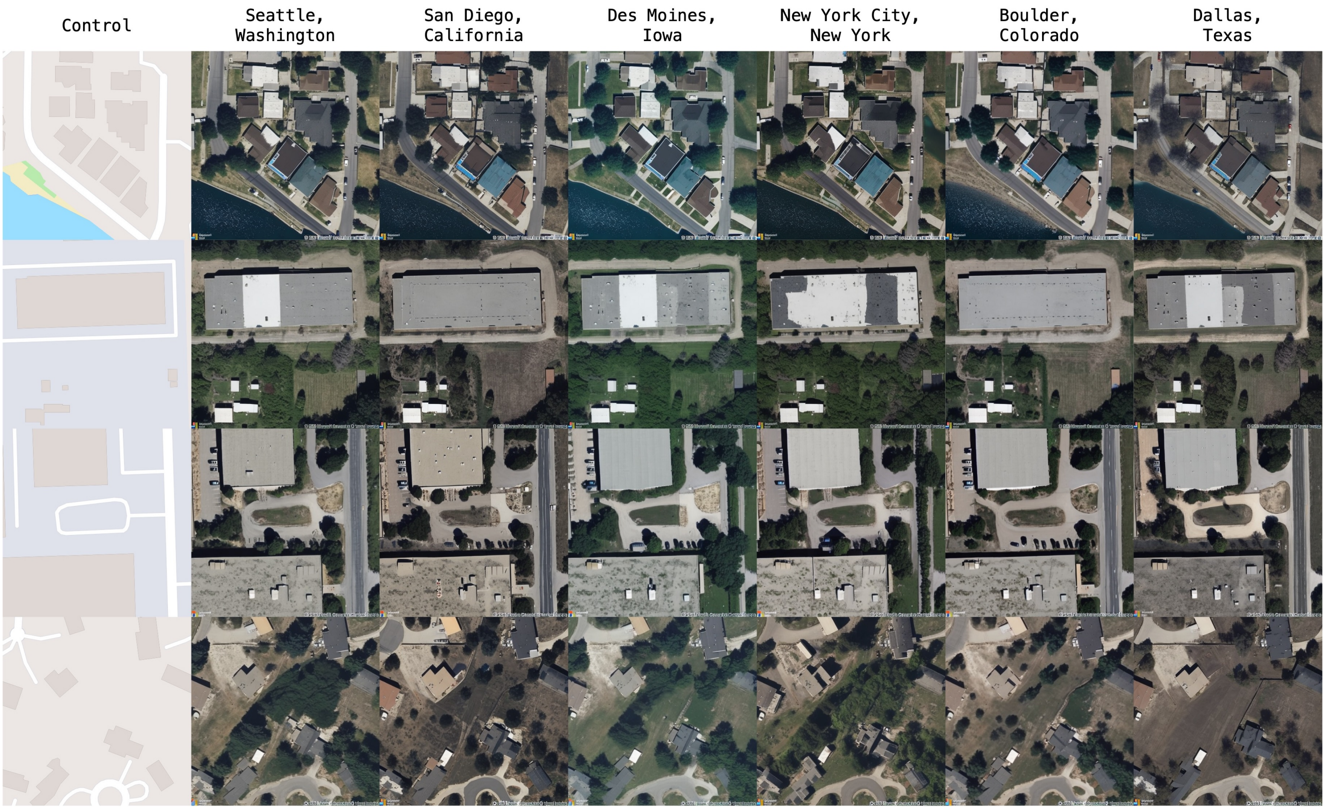

Figure 4. Geo-aware generation. We show four example generations of satellite images using six different geographic locations. We use the same OSM control and random seed without specifying any textual prompt.

Method	Control	Location	FID ↓	SSIM ↑	CLIP-Score ↑
GeoSynth	-	-	13.55	0.237	0.287
GeoSynth	-	✓	12.01 (**+1.54**)	0.264 (**+0.027**)	0.288 (**+0.001**)
GeoSynth	Canny Edge	-	15.35	0.350	0.291
GeoSynth	Canny Edge	✓	13.92 (**+1.43**)	0.361 (**+0.011**)	0.289 (-0.002)
GeoSynth	SAM Mask	-	12.29	0.335	0.297
GeoSynth	SAM Mask	✓	12.04 (**+0.25**)	0.346 (**+0.011**)	0.290 (-0.007)
GeoSynth	OSM	-	12.97	0.274	0.298
GeoSynth	OSM	✓	11.90 (**+1.07**)	0.291 (**+0.017**)	0.303 (**+0.005**)

Table 1. Incorporating geographic location as an additional condition results in higher FID and SSIM scores. For each of these experiments, we incorporated text prompts during the training.

CLIP-score [31] to measure the similarity between synthesized images and corresponding text prompts.

6. Results and Discussion

Geo-aware synthesis. In Figure 4, we demonstrate four example syntheses from varying geographic locations. We use the same control OSM image and random seed for each example. Additionally, no prompt was provided for synthesizing the images. It is observed that our models have learned high-level semantics of various geographic locations across the USA. This is confirmed by qualitatively examining the synthesized images, where Iowa is characterized by more greenery while California has more desert-like features. Geographic locations in and around New York tend to produce satellite images with heavy urban development. Table 1 presents a quantitative evaluation

Figure 5. Synthesis performance of GeoSynth when using various layout controls and text prompts.

Method	Control	Text	FID ↓	SSIM ↑	CLIP-Score ↑
GeoSynth	-	-	16.11	0.199	0.207
GeoSynth	-	✓	13.55 (**+2.56**)	0.237 (**+0.038**)	0.287 (**+0.080**)
GeoSynth	Canny Edge	-	16.74	0.200	0.274
GeoSynth	Canny Edge	✓	15.35 (**+1.39**)	0.350 (**+0.15**)	0.291 (**+0.017**)
GeoSynth	SAM Mask	-	13.48	0.268	0.262
GeoSynth	SAM Mask	✓	12.29 (**+1.19**)	0.335 (**+0.067**)	0.297 (**+0.035**)
GeoSynth	OSM	-	12.70	0.273	0.269
GeoSynth	OSM	✓	12.97 (-0.27)	0.274 (**+0.001**)	0.298 (**+0.029**)

Table 2. Text-guided training improves the quality and diversity of synthesis.

of the models when incorporating geographic location as an additional condition. Adding geographic location improves the ability of the models to create satellite images that look more realistic, as confirmed by better FID and SSIM scores. However, the CLIP-score shows little to no improvement, which is expected since this score measures the similarity between image and the corresponding text prompt. Text-only GeoSynth achieves a high FID and CLIP-score while receiving a poor SSIM score. This indicates that the model can produce semantically meaningful images that differ from the original ground-truth distribution at the pixel level.

Control Images. In Figure 5, we show satellite images synthesized using three controls: Canny edge image, SAM mask, and OSM image. We show result synthesis when using various challenging prompts. It is noticed that the model using Canny edge as a control generates the most realistic-looking satellite images, as confirmed by the high SSIM score. However, it is incapable of regulating the style of the satellite image as given in the prompt. Although SAM mask achieves the highest scores on average over all the metrics, it fails to produce visually aesthetic-looking satellite images. This happens due to the over-segmented masks produced by SAM when applied to satellite images. As proven by the highest FID and CLIP-score, OSM imagery as control produces the most semantically meaningful

Figure 6. Editing. We show three example generations of satellite images using edited versions of a control OSM image. We use the same random seed without specifying any textual prompt.

Figure 7. Example synthesis of GeoSynth using out-of-domain control image.

satellite images. Furthermore, the model shows a good zero-shot synthesis capability. It can effectively control the style of satellite images according to the prompt. As depicted in Figure 6, our model can be utilized to edit satellite images by providing edited copies of OSM imagery. This is possible by using the same textual prompt and random seed.

Importance of text. We experimented to determine the significance of text guidance in satellite image synthesis, as shown in Table 2. The models were trained both with and without text prompts, disregarding geographic location. In instances where we trained models without text guidance, we provided an empty string as the text prompt. Our findings indicate that the performance of GeoSynth is poor when it is trained without including any text. When text is not incorporated, the model is incapable of generating diverse images. However, by including text, GeoSynth can produce realistic-looking satellite images. Similar observations are made when models are additionally trained with control images. Across all the metrics, a significant gain is seen in the CLIP-score. This is expected since CLIP-score reflects the similarity between a textual prompt and the corresponding synthesized image.

Zero-shot capabilites. We evaluated zero-shot generalization of our models. Firstly, we show the synthesis performance of our model when using out-of-domain control images. In Figure 7, our model was provided with

Class	CLIP-Confidence ↑
airport	55.41
amusement park	57.90
beach	92.74
botanical garden	56.94
factory	78.78
farmland	91.81
golf course	87.35
harbor	52.27
parking lot	92.46
railway station	70.94

Table 3. CLIP zero-shot classification performance on the synthesized samples generated using GeoSynth with a fixed OSM image as control.

a control image in the form SAM mask. Visually, the model has synthesized realistic-looking satellite images. Similar behavior is observed when the model is provided out-of-domain OSM or Canny edge image. Next, we assessed the performance of our model in generating a variety of concepts while using a fixed control image. We selected ten land-use categories and synthesized 50 images for each category, using fixed OSM imagery. We provided the names of these categories in the textual prompt and maintained a consistent random seed throughout the generation process. We employed CLIP's zero-shot classification pipeline to classify each synthesized image into a set of binary classes. For each category, we determined whether the generated images belonged to that category or not. In Table 3, we report the average confidence value of CLIP for each category. A higher score indicates that CLIP classified a synthesized image into a given land-use category with high confidence. Our results indicate that the generated images effectively represent the specified land-use categories. Across all the land-use categories, our model was able to achieve an average CLIP-confidence of 73.66, which indicated an image generated using our model depicted the correct land-use category 73.66% times on average. Lastly, we evaluated the model's performance in synthesizing images of categories specified in the UCMerced dataset [52]. UCMerced contains satellite images at 0.3m resolution across 21 land-use categories. Figure 8 depicts the zero-shot synthesis capability of text-only GeoSynth model. Table 4 demonstrates the class-wise performance of our models on the UCMerced dataset. Overall, the models perform well in certain categories such as beaches, buildings, etc. On the other hand, they perform poorly in categories such as storage tanks, airplanes, etc. We observe that the performance of GeoSynth with Canny Edge image or SAM mask as layout control on UCMerced depends highly on the quality of the layout image itself. Most often,

Figure 8. Generated samples from our GeoSynth model, without layout control, on UCMerced classes.

Class	GeoSynth (No Control)		GeoSynth (Canny Edge)			GeoSynth (SAM Mask)		
	FID ↓	CLIP-Score ↑	FID ↓	SSIM ↑	CLIP-Score ↑	FID ↓	SSIM ↑	CLIP-Score ↑
agricultural	33.95	0.261	29.56	0.056	0.245	38.17	0.013	0.246
airplane	32.44	0.245	50.57	0.090	0.227	70.19	0.087	0.222
baseballdiamond	22.57	0.307	42.97	0.089	0.276	48 .81	0.185	0.280
beach	32.94	0.237	39.07	0.271	0.214	43.73	0.200	0.216
buildings	26.97	0.245	33.63	0.120	0.253	56.23	0.084	0.247
chaparral	55.46	0.222	46.04	0.123	0.216	51.72	0.088	0.206
denseresidential	26.80	0.267	26.15	0.113	0.261	51.39	0.058	0.252
forest	18.48	0.265	32.35	0.091	0.245	37.24	0.045	0.241
freeway	22.86	0.258	36.61	0.086	0.265	65.60	0.071	0.243
golfcourse	23.88	0.274	39.66	0.209	0.261	44.18	0.177	0.262
harbor	69.26	0.203	29.32	0.137	0.224	38.90	0.074	0.230
intersection	27.25	0.271	22.78	0.152	0.272	37.73	0.101	0.275
mediumresidential	20.13	0.254	35.98	0.109	0.252	69.79	0.055	0.238
mobilehomepark	35.22	0.287	50.79	0.109	0.257	61.74	0.048	0.243
overpass	21.61	0.262	47.40	0.078	0.263	65.61	0.080	0.247
parkinglot	36.37	0.278	34.70	0.097	0.256	49.66	0.060	0.243
river	21.63	0.237	53.68	0.119	0.223	62.66	0.083	0.219
runway	45.78	0.230	53.64	0.100	0.214	57.38	0.084	0.214
sparseresidential	36.49	0.262	36.95	0.125	0.240	48.16	0.104	0.238
storagetanks	33.12	0.285	55.71	0.125	0.233	64.89	0.103	0.243
tenniscourt	23.37	0.296	57.01	0.101	0.238	50.90	0.097	0.257

Table 4. The performance of zero-shot synthesis of our models on UCMerced categories.

SAM produces undersegmented images when applied on UCMerced. On the other hand, the Canny algorithm produces a lot of false edges. In the future, it is possible to improve the overall performance by finetuning the models on additional datasets.

7. Conclusions

Text-to-image models have exhibited impressive performance and have been widely used in various end-user applications. However, there has been little to no research conducted on this topic in the field of remote sensing. This lack of adoption of such models in remote sensing represents a missed opportunity for building innovative applications. Therefore, we aim to inspire the remote sensing community and promote future research directions in conditional satellite image synthesis through this work. While we leave potential applications of our framework as future work, we believe that urban planners will benefit the most from it. One could imagine using our framework for automatic digital twin generation, urban growth simulation, and city planning. To encapsulate, we proposed GeoSynth, a suite of models capable of synthesizing realistic-looking satellite images while allowing personalization through text prompts. It uses spatial layout from input control images to guide the synthesis process. Additionally, our model incorporates geographic location as a condition that improves the synthesis quality by considering a region's geographical features. We hope GeoSynth represents the first step towards a global geography-aware synthesis model.

References

[1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 2

[2] 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. 1, 2

[3] Tim Brooks, Bill Peebles, Connor Homes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Wing Yin Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1, 2

[4] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 2

[5] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5343–5353, 2024. 2

[6] Ning Chen, Jun Yue, Leyuan Fang, and Shaobo Xia. Spectraldiff: A generative framework for hyperspectral image classification with diffusion models. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 2

[7] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[8] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery, 2023. 1

[9] Aayush Dhakal, Adeel Ahmad, Subash Khanal, Srikumar Sastry, and Nathan Jacobs. Sat2cap: Mapping fine-grained textual descriptions from satellite images, 2023. 1

[10] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[11] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 2

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[14] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023. 1, 2

[15] Yutong He, Dingjie Wang, Nicholas Lai, William Zhang, Chenlin Meng, Marshall Burke, David Lobell, and Stefano Ermon. Spatial-temporal super-resolution of satellite imagery via conditional pixel synthesis. *Advances in Neural Information Processing Systems*, 34:27903–27915, 2021. 2

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 2

[17] 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. 4

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2, 3

[19] Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Weiming Dong, and Changsheng Xu. Diffstyler: Controllable dual diffusion for text-driven image stylization. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 2

[20] Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi, Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation models for generalist geospatial artificial intelligence, 2023. 1

[21] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 1, 2

[22] Subash Khanal, Srikumar Sastry, Aayush Dhakal, and Nathan Jacobs. Learning tri-modal embeddings for zero-shot soundscape mapping, 2023. 1

[23] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. *arXiv preprint arXiv:2312.03606*, 2023. 2, 3

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[25] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 2023. 2, 4

[26] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 2

[27] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[29] Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14267–14276, 2023. 2

[30] Yotam Nitzan, Michaël Gharbi, Richard Zhang, Taesung Park, Jun-Yan Zhu, Daniel Cohen-Or, and Eli Shechtman. Domain expansion of image generators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15933–15942, 2023. 2

[31] 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. 3, 5

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1, 2

[33] Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning, 2023. 1

[34] Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Mission critical - satellite data is a distinct modality in machine learning. *ArXiv*, abs/2402.01444, 2024. 2

[35] 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. 1, 2, 3

[36] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[37] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 2

[38] 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. 1, 2

[39] João Daniel Silva, João Magalhães, Devis Tuia, and Bruno Martins. Large language models for captioning and retrieving remote sensing images, 2024. 1

[40] 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. 1, 2, 3

[41] Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, et al. Styledrop: Text-to-image synthesis of any style. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[42] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries, 2024. 1

[43] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2

[44] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[45] Vicente Vivanco, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. 2023. 1

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[47] Yi Xiao, Qiangqiang Yuan, Kui Jiang, Jiang He, Xianyu Jin, and Liangpei Zhang. Ediffsr: An efficient diffusion probabilistic model for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 2

[48] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023. 2

[49] Jianjin Xu, Saman Motamed, Praneetha Vaddamanu, Chen Henry Wu, Christian Haene, Jean-Charles Bazin, and Fernando De la Torre. Personalized face inpainting with diffusion models by parallel visual attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5432–5442, 2024. 2

[50] Yonghao Xu, Weikang Yu, Pedram Ghamisi, Michael Kopp, and Sepp Hochreiter. Txt2img-mhn: Remote sensing image

generation from text using modern hopfield networks. *IEEE Transactions on Image Processing*, 2023. 2

[51] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14256–14266, 2023. 2

[52] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010. 7

[53] Angelos Zavras, Dimitrios Michail, Begüm Demir, and Ioannis Papoutsis. Mind the modality gap: Towards a remote sensing vision-language model via cross-modal alignment. *arXiv preprint arXiv:2402.09816*, 2024. 1

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3, 4

[55] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain, 2024. 1

[56] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2