Map Diffusion - Text Promptable Map Generation Diffusion Model

Marcin Przymus
mprzymus@kraina.ai
Kraina AI - Geospatial & Mobility Research Group
Department of Artificial Intelligence
Wrocław University of Science and Technology
Wrocław, Poland

Piotr Szymański
piotr.szymanski@pwr.edu.pl
Kraina AI - Geospatial & Mobility Research Group
Department of Artificial Intelligence
Wrocław University of Science and Technology
Wrocław, Poland

Figure 1: Examples of map tiles generated using the Map Diffusion model conditioned on different prompts.

ABSTRACT

This paper introduces a novel text promptable map generation model, leveraging recent advancements in generative models. Promptable map generation has broad applications, democratizing access to geographic data, enhancing decision-making, improving communication, and enabling customization. Map Diffusion generates maps based on textual descriptions, allowing users to describe a region, and the model generates a corresponding map. We conduct a comprehensive review of related work, highlighting the unique contributions of our model. We also provide insights into dataset creation, model architecture, training procedures, and experimental results. This research marks a significant step in harnessing generative models for map generation, opening doors for future exploration in this field.

CCS CONCEPTS

• **Applied computing** → **Environmental sciences**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

map generation, diffusion models, generative models, text to map generation

UrbanAI '23, November 13, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0362-1/23/11…$15.00
https://doi.org/10.1145/3615900.3628787

ACM Reference Format:
Marcin Przymus and Piotr Szymański. 2023. Map Diffusion - Text Promptable Map Generation Diffusion Model. In *1st ACM SIGSPATIAL International Workshop on Advances in Urban-AI (UrbanAI '23), November 13, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3615900.3628787

1 INTRODUCTION

In recent years, we have witnessed the dynamic development of generative models, especially those that allow for generating images conditioned on textual descriptions, known as prompts. They have found applications in various fields of activity, ranging from simplifying repetitive graphic tasks to significantly enhancing users' creative expression and even greatly facilitating interpersonal communication through rapid visualizations of thoughts that previously required effort on the part of the sender to create and then understand by the recipient. In this paper we propose the - to our knowledge first - text promptable map generation model, that builds on top of that progress.

Promptable map generation is important because it democratizes access to geographic data, enhances decision-making, improves communication, allows for customization. It has the potential to transform how individuals and organizations interact with and use geographic information. Makes it easier for people who may not have expertise in GIS or mapping software to create maps. It can enhance communication by enabling people to easily create maps that illustrate their ideas, data, or findings, whether for presentations, reports, or educational materials. Generative models showed their robustness in many tasks and numerous domains. It is reasonable assumption that they can be also useful in this area. The use of text-to-image models such as Stable Diffusion [18] creates high-quality images based on description. Similarly our model could be useful in how users work with maps.

In this paper, we provide **Map Diffusion**[1], a model to generate maps based on textual descriptions (prompts). Specifically, given a description of particular region *d*, the map of region matching this description should be generated. We start with Section 2 which contains a related work review to find advantages and limitations of existing map generation methods and back our claim that the presented model is the first of its kind. We then - in Section 3 - discuss the creation of a dataset[2] containing map tiles and prompts describing them and release the generated dataset publically for future training and benchmark uses. We then describe the model and its training process in Section 4. Section 5 follows with experiments, exploration, and evaluation of the model. Finally we conclude our work in Section 6 and deliberate on potential avenues for future research.

2 RELATED WORK

In this section, we discuss the state of the art with respect to:

- map generation models - works concerning usage of generative image models in geospatial problems,
- natural language geospatial datasets - works tackling and issue of creating geospatial dataset with textual modality.

2.1 Map generation models

GANs for Urban Design (2021). The work presents method of generating urban blocks based on image containing road network and building shapes. The author considers as essential shapes of buildings, their proportions and relationships between them. The architecture of used model is pix2pix. The issue of style transfer was addressed - paper contains evaluation of model trained on one city and tested on another. Additionally, there was resnet [6] based classifier trained on real data to classify city of given block. Classifier was evaluated on synthetic data.

GANmapper: geographical data translation (2022) [25]. In this work translating different types of map data to generate building footprints. The experiments were conducted with black and white road networks, coloured road diagrams (CRHD) and land use mask as input. This work uses tiles as input and 4 unique zoom levels are evaluated (14, 15, 16 and 17). The proposed model is based on pix2pix. The work raised the issue of stitching adjacent tiles into one image. To evaluate results Fréchet inception distance (FID) [7] and Mean Intersection over Union (mIoU) was used. The results showed that roads are much more powerful source of information to use than land use mask. The zoom results showed general rule the bigger zoom, the better FID is. Authors believe that for cities with larger buildings footprint the best trade-off between image quality and size of spatial context is provided by zoom 15 or 16. The study shows that model is capable of keeping spatial consistency on stitched outputs.

InstantCITY: Synthesising morphologically accurate geospatial data for urban form analysis, transfer, and quality control (2023) [26]. This works follows above one direction and develop it with model generating higher resolution images. To achieve this, pix2pixHD model [23] was used. Model use two generators: G1 for lower resolution G2 for higher (1024x1024). Images are judged by 3 discriminators to achieve well quality in context of both details and structural consistency.

Building layout generation using site-embedded GAN model [9]. This work proposed method using not only image data as input. It is focused on generating only building layouts (with its height) in scale of census blocks (CB). Work compares result of including in input only boundaries, boundaries and conditional vector or boundaries, conditional vector and buildings centroids. The extra tabular data gives significant improvement of generation.

Street Layout Design via Conditional Adversarial Learning (2023). [27] This work presents application of conditional GANs to street network layout generation. Authors focused on roads tagged on OSM as highway. Generator is fed by terrain representation from autoencoder model trained on three single channel images for elevation, population density and land use. Based on generator's output they build street network graph.

Diffusion models are a relatively new type of image generation model, but they have quickly become the most powerful type of image generation model available. They achieve high image quality, controllability and versatility. Multiple methods [4] [8] [20] were used successfully to adopt diffusion models to certain domains. This shows their high capability to adopt to miscellaneous problems.

High-Resolution Image Synthesis with Latent Diffusion Models (2021) [18]. This work introduced model known as Stable Diffusion. Authors proposed latent diffusion models what means that denoising U-net network works in latent space. Despite the fact that this approach reduces resources needed to train a model, they achieved competitive results for many tasks. Article introduces a conditioning mechanism using cross-attention which is not limited by conditioning modality. This conditioning can work with texts, images, semantic maps and representations from neural networks.

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding(2022) [21]. The work present another diffusion model with results slightly better than Stable Diffusion. The interesting contribution is they they explored usage of large language models instead of multimodal models like CLIP [16]. Moreover, they reported human evaluators preferred model with T5 [17] as encoder instead of CLIP. High resolution (1024x1024) was achieved by generating low resolution image (64x64) and then twice applying it to super-resolution diffusion model.

LoRA: Low-Rank Adaptation of Large Language Models (2021) [8]. This work introduces a method for fine tuning large models. The goal of this method is to reduce number of trainable parameters and the training time. LoRA works by freezing the pre-trained model weights and injecting trainable rank decomposition matrices into each layer of the transformer architecture. This reduces number of trainable parameters and the storage space required for the updated model. This method was successfully used with diffusion models, e.g. with Stable Diffusion. The most standard approach is to apply LoRA on cross attention layers which attends between image and prompts in denoising network.

[1] https://huggingface.co/kraina/map_diffusion
[2] https://huggingface.co/datasets/kraina/text2tile

2.2 Natural language geospatial datasets

Mentioned works didn't combine natural language prompting and geospatial data. Due to this fact there are no datasets combining those modalities mentioned in above sections. To address this lap, following works are presented:

RSVQA meets BigEarthNet: a new, large-scale, visual question answering dataset (2021) [12]. The work presents RSVQAxBEN dataset which is made up of photos and question-answer pairs. Based on existing labels two types of question are created - yes/no about presence of label and land cover questions. 80.7% of questions are yes/no questions. Totally, dataset contains 10,659 images and 955,664 questions. In context of this work this dataset seems to be inaplicable due to image format and no textual answer, however it shows that textual information about spatial region can be created allegorically - as questions in this example.

Deep Semantic Understanding of High Resolution Remote Sensing Image(2016) [15]. This work introduced interesting dataset. It contains 2100 images which were divided into 21 groups. Each image is annotated with 5 corresponding sentences. Despite the fact dataset contains 10,500 captions, only 2023 of them are unique. The most common seems to be generic - "There is a piece of farmland/cropland" or other with same meaning. Sentences present only once seems to be complicated, however not all, like "Some buildings here". The dataset shows simple captions pointing out objects present in images, however cannot be used directly because of images style. Furthermore, this dataset would be much too small for our work.

Above, we discussed generative models, with particular emphasis of those used in map application, were presented. Majority of them focuses on generating building blocks. Some other works focuses on roads. In roads topic is present 2 approaches - generating them as a graph [1] or firstly generating image [27] [5]. We note that all of those models are image promptable models, not text promptable. Image context modality (inpainting task) is highly explored area - all models are capable of using it. However, diffusion models are also capable of using text. Authors of [18] mention their model readiness of taking embeddings of any type of data. Our literature review shows lack of textual description usage for map generation task. In task of text-to-image generation their robustness denoising diffusion models has shown. Reviewed models achieve similarly good results and all of them should be capable to handle map data. Due to easy availability including a few pretrained checkpoints and interesting possibility of finetuning [20] [4] Stable Diffusion model from [18] seems to be convenient framework to be used in this paper.

3 DATA

As with text promptable models, also there are no usable dataset for promptable map generation, i. e. datasets that would include a prompt rich enough to allow a good level of control over the conditioned model while also remaining somewhat human readable. In this section we discuss and prepare such a dataset.

3.1 Collected area

For data collection set of 62 cities from 4 continents was chosen. Those cities were used previously in literature and misses cities with hard to define boundaries [24]. Totally, it contains 164,662 tiles. Based on GanMapper[25] zoom level 16 was used which is trade-off between details and wider context of terrain. Dataset contains tags information which are used to generate captions and raster images of tiles. Additionally, binary buildings mask were generated for this dataset.

Small amount of tiles (2%) misses any tag. This is not large amount and do not have to be removed because such tiles might be useful to learn model general map style. The median is equal to 11, mean 17.71. It means that sentence with reasonable length usually contains great part of information included in tags occurrences. Tiles with less than 9 tags represents 42.70% of all samples. To check how empty are tiles with small number of miscellaneous tags, total number of objects in tiles from this subset needs to be investigated. 24.73% of them contains only one object. Median is 3 and mean 3.47. It shows that even less populated areas may contain some interesting data.

4 least popular tags are marginal in terms of the number of occurrences. However, water shares token "water" in CLIP tokenizer used by stable diffusion. Objects with healthcare key usually occurs with amenity and building tags and shares same values (like "doctor"), so it is seen more often than number of keys suggest. Military and aeroway are connected with specific areas such airports and military ares. Correlation between keys quantities are shown in figure 2. Alone key values does not show many interesting relation between them. Water and waterway tags quantity in tiles does not correlate with anything expect themselves and natural. They have weak negative correlation with shops and offices.This is the only negative correlation in dataset, meaning that generally more populated tiles are more populated with any tags. This phenomena may be related with crowd based tagging mechanism in OSM. Figure shows some intuitive relation, like between "natural" and "water".

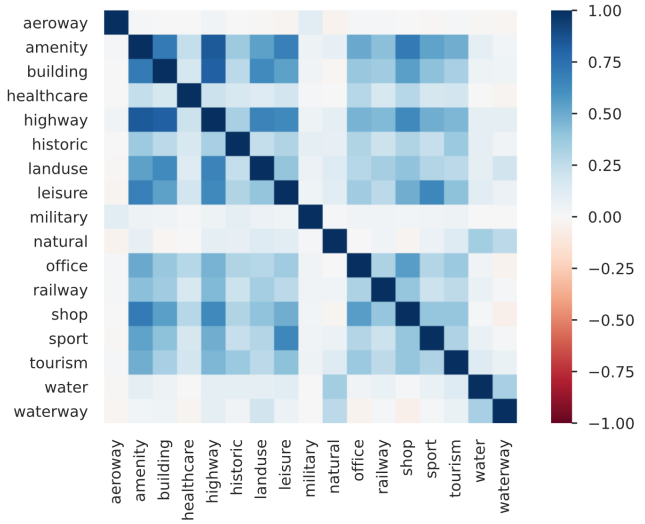

Figure 2: Correlation between tags keys

Sentences
OSM from Chicago, United States, North America of hamlet area containing: 487 buildings, 18 highway services, 3 landuse residentials.
OSM from Łódź, Poland, Europe of quarter area containing: 1 highway unclassified, 8 highway living streets, 1 shop hairdresser.
OSM from Madrid, Spain, Europe of suburb area containing: 9 highway tracks, 1 highway service, 1 highway primary.
OSM from Moscow, Russia, Europe of area containing: 1 leisure sports centre, 13 highway footways, 4 building industrials.
OSM from Gdańsk, Poland, Europe of unknown area containing: 1 natural bay.
OSM from Łódź, Poland, Europe of hamlet area containing: 35 building houses, 2 highway services, 1 landuse construction.
OSM from Madrid, Spain, Europe of neighbourhood area containing: 4 highway paths, 20 highway tracks, 1 landuse forest.

Table 1: Examples of caption generation for 3 features.

The most popular tag in whole dataset is pair "highway: service" and "building: yes". The most popular tags are dominated with roads related pairs - half of 20 most common pairs is traffic related. In group of most popular are also residential related tags (landuse residential, parking) and green areas related such as parks and forests. This shows that model trained on this dataset should not be biased on city centres. Average tile contains more than 40 buildings. The most often value for building is simply "yes" which is related with over 30 buildings on average. Footways are the most numerous ways type in dataset, despite the fact that more tiles are populated with service and residatial ways.

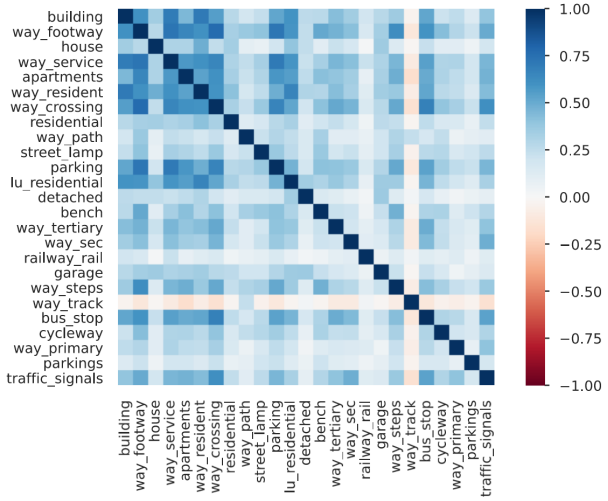

Figure 3: Correlation between top 25 most popular tags

Most popular tags tend to highly correlate. Exceptions are not urbanized tags such as forests and farmlands. Most high-density areas are residential. Tiles with number of buildings over mean almost always (94%) contains residential roads and usually are tagged with residential landuse (83%). In such areas parkings and bus stops are popular similarly. Around 60% of such tiles contains them. 40% of such tiles contains parks. This value depends on city, e. g. in Los Angeles it drops to 28%. Well populated with buildings tiles often contains also schools, playgrounds, restaurants, cafes and parks. Less building populated residential areas more often lacks of schools and restaurants, however this change does not affect playgrounds. If number of buildings is relatively low (less than 10), most common type of road is path. Keys popularity shows two main groups in data: green terrains such as forests or farms and highly urbanized areas rich with buildings and amenities such as restaurants, schools, footways and parking spaces. Almost half of dataset are areas without any buildings. They are much worse communicated, however half of them has some highways. 20% of them are tagged as nature reserve. It's important to underlay that parks usually are not big enough to take all tile so tiles with park usually contains typically urban objects. Majority of those tiles has number of buildings over the mean. Data contains to types of industrial areas. First one is big separated factories. They are overpopulated (in comparison with rest of data) with service roads, bus stops, parking spaces and industrial buildings. Second one contains some industrial building, however such areas has much more single-family houses than rest of tiles and are more often tagged as residential areas.

3.2 Captions

Captions are automatically generated from OSM tags data and the result is similar to those from UMC dataset[15]. Key-value pairs occurrences was counted for each tile. OSM contains great number of tags which can be added manually by users. Due to this fact tagging might be inconsistent depending on people who added information to OSM. To avoid this issue a good subset of used tags must be chosen. In this work subset proposed in hex2vec [24] was used. Hex2vec does not cover transportation so extra set of tags for transportation data was needed. To address this issue tags with keys railway and highway present in OSM wiki were chosen to cover transportation. The procedure of downloading OSM data was implemented with srai library. Using those tags, method of building representation for small hexes of cities was proposed.

To generate sentence from set of all present in tile tags small number (which is limited by text encoder maximal number of tokens) was sampled randomly. Each sentence is constructed with 10 tags (or less if tile contains less elements). Due to the fact that usually tile contains small number of present tags loss of information at this stage is not big and should not influence on model's behaviour. This number also is small enough not to exceed text encoding model maximal number of tokens. Dropping of some information in caption should be profitable also because redundancy in tag information. Tags such as amenity bench are not explicitly shown in raster tile, however are most often in areas with parks than in whole dataset. This way, dropping of some tags in caption should help model to learn non obvious relations and predict some necessary dependencies in city structure. The idea of hiding some information in order to achieve better subject understanding is popular in literature, e. g. in BERT [2] models. Captions are generated before training and for each tile same caption is used in whole training process due to performance issues, however giving model another version of prompt for tile at each epoch might be profitable to achieve better urban structure understanding.

Sentences starts with "OSM" which is unbiased word by model so it gives opportunity to connect to OSM tile style easier. Other words like map or tile are highly connected with known to model ideas. Next section describes location. This includes city, country and continent. To precise part of city, admin level tag was used. For each tile object with the highest (most precise) admin level was found. Then tag place was used. This process results with values such as *suburbs, neighbourhood, quarter*. Then, chosen pairs are listed in sentences with number of occurrences separated with comma. Sometimes key is simply word "yes" (building: yes). In such scenario word yes is skipped to keep natural appearance of sentence. Examples of generated data are in table 1.

3.3 Image data

For the same tiles as the captions, raster images were downloaded and building masks were created. A custom tile style was created based on the osm-bright style. The most important difference from the original style is the removal of all text information. Diffusion models are not designed to learn text in an image and will not be able to generate tiles with text. In addition, icons such as airport, cafe and supermarket have been removed. They would have given additional information on the map about existing amenities, but the model would have had trouble correctly reproducing their shapes. The tiles were generated using the self hosted tileserver-gl. The tiles have a resolution of 256x256.

To evaluate information about buildings from generated tiles model which separates buildings from rest of tile is needed. To train such model tile buildings mask was created. Those masks mark all object with "building" key. To create masks another style was created. Tiles were generated and saved with same procedure as raster tiles.

Due to lack of sufficient dataset for the intended goal, a new dataset was created. We make the dataset available publically on the HuggingFace website[3]. The dataset contains a large amount of tiles which is sufficient to train the generative model. Each tile has prompt with information about types and amount of objects in tile. The dataset also contains building masks for each tile, which are useful to train the model which extracts the building layer from a generated tile.

[3]https://huggingface.co/datasets/kraina/text2tile

Figure 4: Raster tile and corresponding to it buildings mask.

4 PROPOSED MODEL

We proceed to describe the proposed model, making the first step into the unexplored field of denoising diffusion models application on map-related tasks. We used Stable Diffusion, which is a widely used open model achieving good results in many fields. The checkpoint used as starting one was 1.5. Stable Diffusion 2.0 was considered, which has the same u-net architecture but a bigger text encoder, but performance of both models is similar.

4.0.1 Stable diffusion model. The model used in this work comes from the latent diffusion models family. As shows in figure 5 image x from pixel space is encoded to z before the diffusion process and z obtained after denoising process is decoded to $\widetilde{x}$. The z is representation of x in less dimensional latent space. Conducting diffusion processes in latent space allows to save computational resources in both training and inference mode. To map pixel space

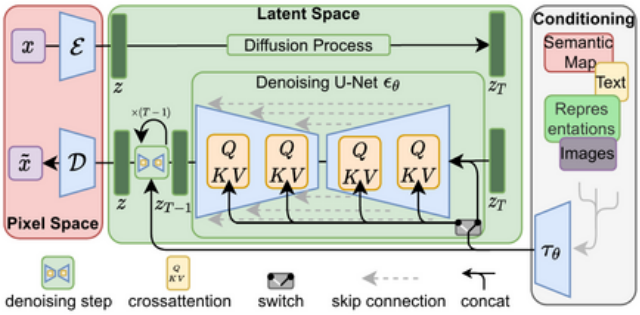

Figure 5: Base model architecture, [18]

into latent space Variational Autoencoder (VAE) [10] is used. As typically, a Gaussian distribution is used. To make latent space of prior distribution, Kullback-Leibler divergence between a prior distribution and $q(z|x)$ is minimized.

To denoise image u-net network is used. U-Net is a convolutional neural network (CNN) architecture that is commonly used in diffusion models. U-Net consists of two main parts: an encoder and a decoder. The encoder is responsible for extracting features from noised input image, while the decoder is responsible for reconstructing the image from the extracted features. Each u-net layer is also given with timestep embedding and conditioning embedding. The encoder and decoder are made up of resnet blocks followed by cross-attention layers. Cross-attention is used due to conditioning purposes. In this work given condition are text embeddings. To encode prompt CLIP model is used. CLIP is large multimodal model trained with images and its captions. The objective of the CLIP training is to learn a latent space in which images and their captions are close together if they are semantically related, and far apart if they are not semantically related. During the training process, random noise is sampled and added to the image.

For further processing of data generated by the model described in section 4 feature extraction model is proposed. The goal of this model is to extract the shapes of objects from the desired layer. Such information can be useful to compare the model's performance with related works which focus on one layer. Such defined problem is a kind of segmentation. Due to the fact that map layers are created from defined colors, smaller models should be capable to handle this task and there is no need to use the most efficient and sophisticated models. If a model trained on authentic data struggles with generated images, the issue will likely originate from the quality of the generated data rather than being a problem inherent to the segmentation model itself. In this paper, we used methods for generated data were investigated:

- Unet - lightweight segmentation model trained from scratch based on [19].
- FCN [13] with pretrained ResNet50 backbone. This old model was used to compare if a bigger (but still relatively small) model can give significant improvement.

We thus presented our model with the architecture of used diffusion model and its interference process, joined with building layer extraction models. We now present their performance in segmentation and generation experiments, implemented with PyTorch [14], PyTorch-lightning [3] and diffusers [22].

5 EXPERIMENTS

5.1 Segmentation models evaluation

Two models were trained for building layer segmentation. The u-net model (0.93 IoU) outperforms FCN (0.83 IoU) on real data.

	Generated tiles	Real tiles
Unet	0.0360	0.0892
FCN	0.0707	0.0896

Table 2: Performance of segmentation models on real tiles - mean building layer size

Table 2 shows that models prediction differs strongly on generated images. A review of generated segmentation masks shows that the trained u-net model is much less sensible in generated images.

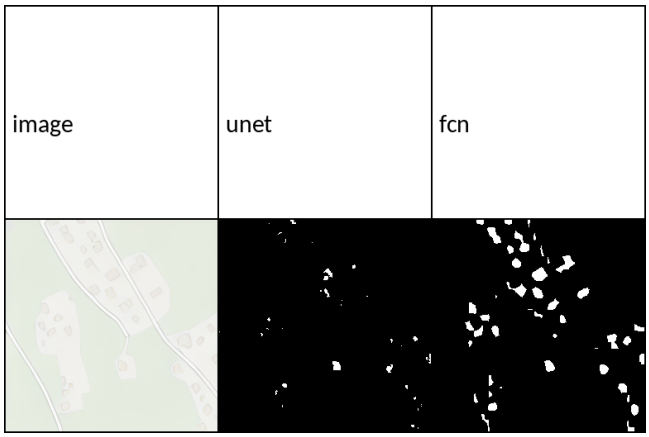

Figure 6: U-net network having low sensitivity

5.2 Quantitative analysis

To measure quality of generated tiles as images FID metric was used. Metric was measured on 10,000 samples. Despite the fact that this metric is not ideal [11], it is widely used including map specific works [25]. Fully trained model achieves 35.01 FID which is quite high, while the quality of images is acceptable.

5.3 Qualitative analysis

5.3.1 City comparison. 5 different cities were compared to show diversity of generated prompts depending on prompted city. Cities were chosen due to their specific characteristics. Wrocław is typical Central Europe distinguished by its heavily branched river. Jerusalem was chosen due to its narrow and winding streets. Lisbon is characterized by its wide spaces between buildings. Stokholm is well developed and diverse city. Los Angeles was used to represent American regular urbanization style. Diversity was shown on example of 5 prompts generated from typical scenarios: residential

part of city, city centre, park, sea coast and green area (with green tags which appears in not urbanized area). Prompts are shown in table 3. It's worth noting that sea coast tile prompt does not include words "sea" nor "water" itself. Given with "beach" and "sand", model knows from training data that water is expected. This effect does not work equally well on all cities, but it is worth noting. Grid of generated with them tiles is presented in figure 7.

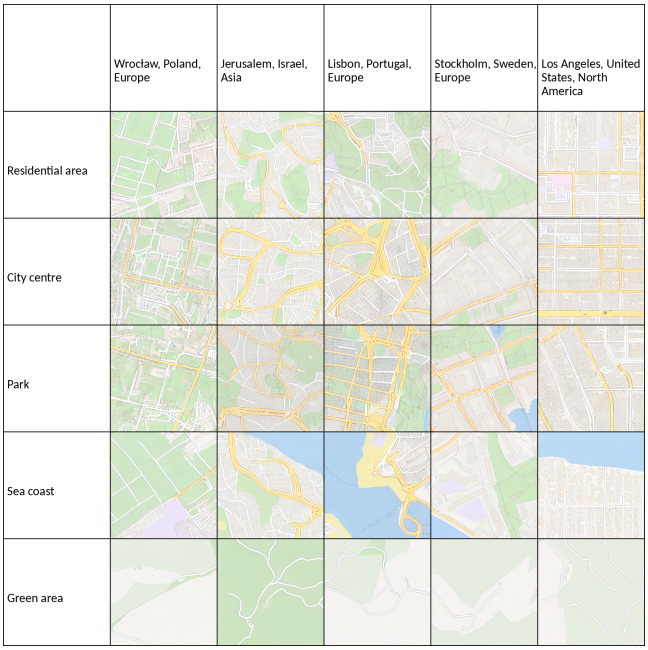

Figure 7: Result of different prompts on cities

The most distinguishable is LA. In first 3 prompts it shows regular, Manhattan-like space management. This one totally ignores park in prompt. This is quite understandable because of small amount of green area in discussed city. Shape of streets changes when sea coast come into tile, however they keeps their well structured character. Despite of tendency to create well structured city centres, model creates almost building free tile when asked. The prompt includes the word building to show that the model understands that it is about lone appearing buildings, not a larger number of them.

For city of Wrocław model generates a lot of green area. Generated tiles for city centre and residential area are not distinguishable easily, due to their dominating green area layer. Model generates some water some water on area when asked, however it tend to be more river like - what fits training data for this city. In Wrocław there is significant area of allotment gardens with many buildings and green background on tiles. Negative prompt "allotment" make model generating more typically urbanized area. In fact, other green land uses in negative prompts, like park or forest achieves similar results. Examples of results with such negative prompt are shown in figure 8. The tiles representing Wrocław exhibit a heightened density of buildings. Nevertheless, the residential region still retains some green space. Importantly, this effect persists even when the negative prompt "park" is substituted with "allotment," indicating that this phenomenon is not solely attributable to the presence of another green area tag not included in negative prompt. This effect remains even if more green tags are added to negative prompt and is characteristic for Wrocław. It is worth noting that regardless of the city in prompt usually tiles keeps similar structure. However, lack of green area layer takes away the hilly character of residential and park tile of Jerusalem.

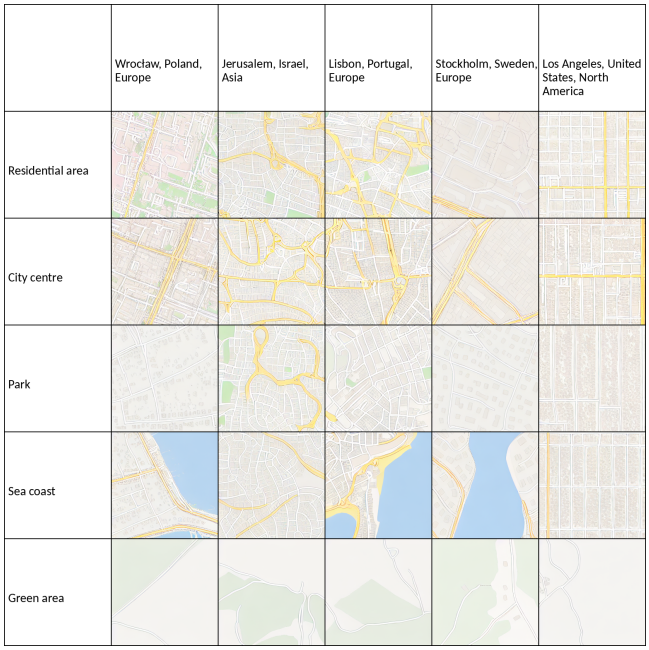

Figure 8: Result of different prompts on cities with negative prompt "park"

Jerusalem's tiles mirror the essence of a city nestled amidst hills. Numerous streets meander as if they were formed within a mountainous landscape. The city center's tiles, devoid of an green expanse, lack this distinctive quality. Not all sections of the city possess such distinctiveness, as can be observed in areas like the Bukharan Quarter. Model could learn such structure from that quarter.

Lisbon is the hardest from discussed cities. Model generates too much chaotic tiles. Model tends to generate main streets more often than in other cities and make then wide. In many tiles there is a problem with logical network of roads - some of them goes nowhere. On the other hand, green space and waters are handled well.

Stockholm tiles have regular structure of roads and large single buildings instead of many tiny packed closed together. It does match real character of city. Shapes of some buildings is not good enough, however it allows to get an imagination of area based on given tile.

5.3.2 Admin level. Prompts uses different area type generated as explained in 3.2. Using residential area prompt from table 3 and Stockholm as city, influence of this information was investigated. As shows figure 11 difference is not significant. Tiles generated with all prompts are presents same type of urban organization. Model is focused on city location information and object which should be

Area type	Sentence
Residential area	OSM from {city} of suburb area containing: 1 amenity kindergarten, 2 landuse allotmentss , 1 building kindergarten, 2 amenity waste disposals , 1 shop beauty, 2 building retails , 1 landuse commercial, 2 amenity pharmacys , 3 amenity recyclings , 3 sport basketballs .
City centre	OSM from {city} of suburb area containing: 1 building garages, 1 shop pastry, 6 tourism informations , 16 highway residentials , 23 tourism hotels, 3 shop greengrocers , 1 railway proposed, 1 office religion, 1 shop locksmith, 1 shop bicycle.
Park area	OSM from {city} of suburb area containing: 1 amenity events venue, 1 landuse brownfield, 3 highway turning circles , 14 building garages , 3 landuse forests , 2 highway unclassifieds , 17 buildings , 63 building houses , 7 highway services , 1 leisure garden.
Sea cost	OSM from {city} of area containing: 3 highway pedestrians , 1 natural beach, 15 highway residentials , 1 sport basketball, 18 building apartmentss , 17 highway services , 1 landuse residential, 38 building residentials , 1 natural sand, 6 highway tertiarys .
Green area	OSM from {city} of city area containing: 1 highway track, 6 landuse forests , 3 highway footways , 2 landuse farmlands , 1 building.

Table 3: Captions used to generate images from figure 7, "{city}" was replaced with name of area in format *city, country, continent.*

included. It does not pay too much attention to area type word. Can be assumed that the reason is behind quality of this data - along the world this names are not always used consistently, so city and included tags are more informative.

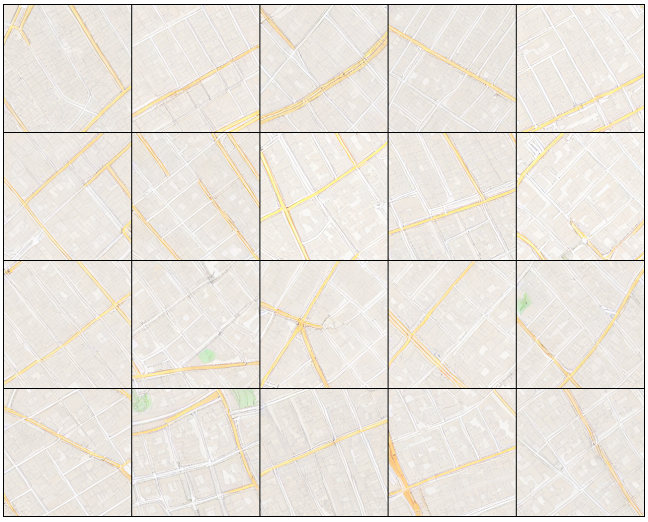

Figure 9: Stockholm and Los Angeles prompt mix.

5.3.3 Mixing city style. Model is capable to mix style of 2 cities. Figure 9 shows result of city centre area prompt given "Los Angeles, Stockholm, Sweden, Europe" as description of location. Generated tiles have dense building development and small amount of green spaces as in LA, however buildings are bigger and road network not so much regular, what follows European urbanization style. Model trained this way can't generate custom mix of styles, e.g. building blocks layer in Stockholm style and road network layer in Los Angeles style. Mix is consequent in many samples for same prompt - model chooses subsets of features of each city, such as building shapes, road network layer or size of green area layer.

Figure 10: Comparison of mixing different USA cities with Poland. First columns shows result for normal city prompt as comparison. Made on park area prompt from 3

Another way to achieve style combination is by blending a city with another continent and country. Style of results is closer to used city than with mixing to cities - continent style is not as clear as city. Continent and country information is weaker than other city and sometimes is not strong enough to change style totally what shows examples of mixture Chicago and Europe in figure 10 where difference is only in small green space in left edge of tile. European style is usually achieved with more area of green spaces (which typically totally lacks in USA) and less quadratic streets. Study shows that model remembers various urbanization styles.

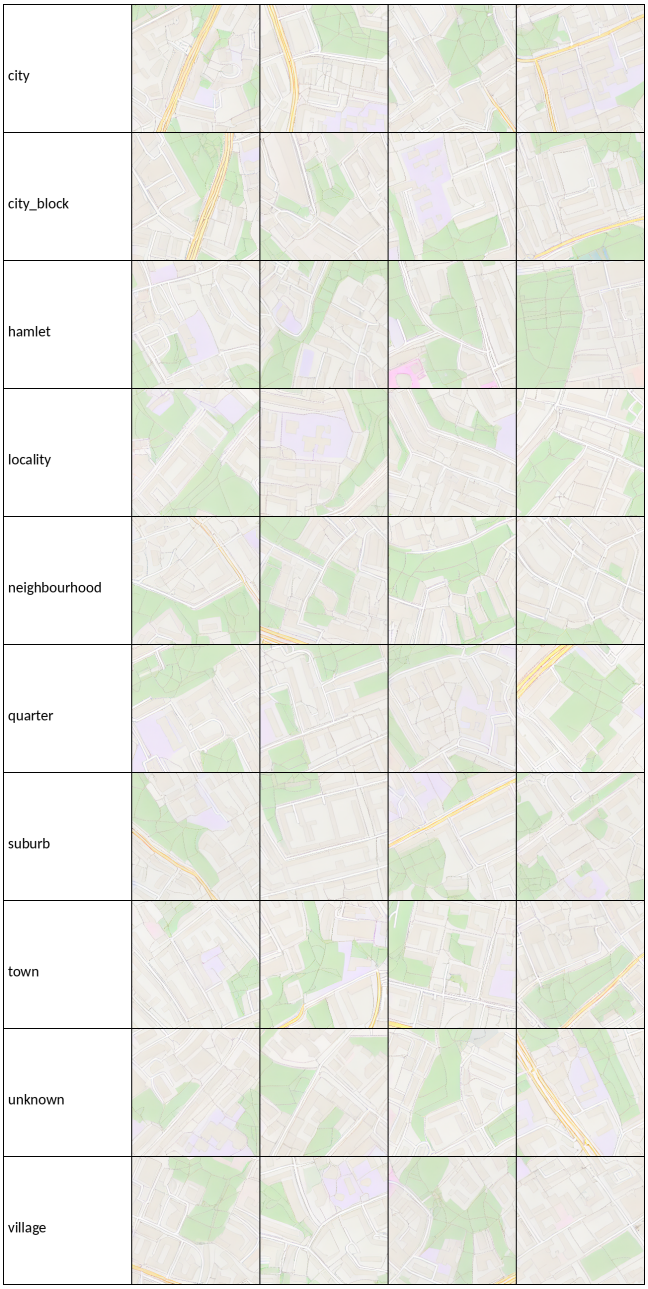

Figure 11: Result of different area types on the same prompt for Stockholm. In each row, there are 4 examples for the same prompt.

We discussed the performance of trained models. The diffusion model generates tiles that are clearly influenced by the information regarding the objects contained within each tile. Individual city styles are visible. The building layer size of generated images is close to the training data. However, some models which work well on true data may be misled by generated data.

6 CONCLUSIONS AND FUTURE GOALS

In this paper, we discussed the map generation problem. We presented the current state of research on this topic and how it lacked data and models for map generation that would use modern, well-performing prompt-to-image diffusion models. We presented a dataset alongside recipes used to create prompts for each raster tile. We conducted experiments and succeeded in training a model that generates reasonable tiles and can be guided by the information about the objects inside tiles and the tile's location. We thus fulfilled the research goal of this paper and obtained a promptable map generation model.

The trained model shows great potential of using diffusion models for map generation, yet it can be improved. First, it could be extended to allow the inpainting task. In this task, the model is given not only text, but also image guidance. An inpainting model would be useful in prototyping small areas since it would consider existing terrain structure. The model prepared is a good starting checkpoint for such training. Second, with larger compute available, the model's generated map area, and context size could be made larger.

REFERENCES

[1] Hang Chu, Daiqing Li, David Acuna, Amlan Kar, Maria Shugrina, Xinkai Wei, Ming-Yu Liu, Antonio Torralba, and Sanja Fidler. 2019. Neural turtle graphics for modeling city road layouts. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4522–4530.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[3] William Falcon and The PyTorch Lightning team. 2019. *PyTorch Lightning*. https://doi.org/10.5281/zenodo.3828935

[4] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. https://doi.org/10.48550/ARXIV.2208.01618

[5] Stefan Hartmann, Michael Weinmann, Raoul Wessel, and Reinhard Klein. 2017. Streetgan: Towards road network synthesis with generative adversarial networks. (2017).

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[9] Feifeng Jiang, Jun Ma, Christopher John Webster, Xiao Li, and Vincent JL Gan. 2023. Building layout generation using site-embedded GAN model. *Automation in Construction* 151 (2023), 104888.

[10] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[11] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2022. The Role of ImageNet Classes in Fr\'echet Inception Distance. *arXiv preprint arXiv:2203.06026* (2022).

[12] Sylvain Lobry, Begüm Demir, and Devis Tuia. 2021. RSVQA meets BigEarthNet: a new, large-scale, visual question answering dataset for remote sensing. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 1218–1221.

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[14] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[15] Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. 2016. Deep semantic understanding of high resolution remote sensing image. In *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*. 1–5.

https://doi.org/10.1109/CITS.2016.7546397

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

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

[20] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. (2022).

[21] 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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

[22] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

[23] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[24] Szymon Woźniak and Piotr Szymański. 2021. Hex2vec: Context-Aware Embedding H3 Hexagons with OpenStreetMap Tags. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*. 61–71.

[25] Abraham Noah Wu and Filip Biljecki. 2022. GANmapper: geographical data translation. *International Journal of Geographical Information Science* 36, 7 (2022), 1394–1422.

[26] Abraham Noah Wu and Filip Biljecki. 2023. InstantCITY: Synthesising morphologically accurate geospatial data for urban form analysis, transfer, and quality control. *ISPRS Journal of Photogrammetry and Remote Sensing* 195 (2023), 90–104.

[27] Lehao Yang, Long Li, Qihao Chen, Jiling Zhang, Tian Feng, and Wei Zhang. 2023. Street Layout Design via Conditional Adversarial Learning. *arXiv preprint arXiv:2305.08186* (2023).