Dataset

Data collection

We collect data from Open Street Map (OSM) (OpenStreetMap contributors 2017). Being an open-source platform, annotations are contributed voluntarily by users. Consequently, OSM frequently contains incomplete markers, especially in sparsely populated areas. For comprehensive and detailed annotations, the dataset is primarily gathered within the Greater London area, at a 1:200 scale. Raw data is acquired using scripts and OSM's application programming interface (API).

The acquired data is formatted in *geojson*, encompassing information about regions, roads, and buildings. Regions encompass the function types (i.e., land use) and their corresponding shapes. Roads comprise road types and layout details. Buildings encompass both building types and shapes. Textual representations are used for function type, road type, and building type. Geometric attributes like region shape, road layout, and building layout are all depicted using sequences of latitude and longitude coordinates.

In accordance with OSM descriptions, we choose 12 function types that are commonly used in cities,as shown in Fig. 1. Finally, we collected more than 147K regions covering 1,300 km^2 as shown in Tab. 1.

Data Processing

Initial data processing involves noise removal. Concerning regions, we merge contiguous ones sharing the same function type, demarcated by roads, while filtering out diminutive regions. Overlapping roads and buildings are removed; redundant attributes like ID, name, address, *etc*., are eliminated. Next, we render our data from geographic format to image format. Because the pre-trained CLIP learns from image-text pairs that are publicly available on the Internet. We render our data using the style of urban layout images on the web (Radford et al. 2021). The urban layout is represented through layered maps, encompassing region layer, road layout and building layout layers. The region layer is rendered with region shape and function type. For the road layout layer, we first categorize roads according to the glossary (for Europe et al. 2010) and then render the road layout with different styles and widths based on the road types. In the building layout layer, building footprints are visualized from an aerial perspective, accurately depicting building shapes. Finally, we create the textural layout description for each region. The layout description describes the urban layout containing function type, road type and building type that this region occupies. For example, a residential region with pedestrian roads and house buildings and residential buildings. Within the layout description, the function type establishes the overarching features. For example, an industrial region typically has regular layouts, while a residential region tends to have irregular and dense layouts. And the road type and building type provide specific features. For example, primary roads are denoted by wide, orange lines, while footways appear as narrow, gray paths. The residential buildings have small shapes, while the commercial building has large shapes. Above all, we obtain a dataset of image-text pairs for our text-driven urban layout regeneration.

Other cities

To assess the generalizability of our approach, we additionally gather urban layout data from Paris and Shanghai. The data collection procedure aligns with the aforementioned introduction. We fine-tune our method and successfully accomplish the text-driven urban layout regeneration in these cities.

Baselines

To the best of our knowledge, our work is the first to handle text-driven urban layout regeneration. We first compare our method with state-of-the-art (SOTA) traditional city modeling methods. (1) CityEngine (Parish and Müller 2001) should carefully design the rule parameters, such as road numbers, road widths, and road template, to generate the road layouts that meet the target as closely as possible. (2) IPSM (Chen et al. 2008) should adjust the tensor field to achieve the target road layout. (3) UrbanBrush (Benes et al. 2021) employs the brush with parameters, such as impact region, population, to create the road layout. Finally, we should manually merge the generated urban layout with the surrounding real urban layout. As we can render the target region with the surrounding context in a bird-view image, we also conduct a comparative analysis of our approach against SOTA text-driven image synthesis techniques. (4) TDANet (Zhang et al. 2020) is a text-guided dual attention network for image inpainting. (5) BD (Avrahami, Lischinski, and Fried 2022) utilizes text to edit the content of the target region. (6)StyleCLIP (Patashnik et al. 2021) is a text-to-image GAN model and can inpaint the content conditioned on the text prompts. (7) SDv2 (Rombach et al. 2022) is a latent text-to-image diffusion model and can inpaint the content conditioned on the text prompts. We modify these models to suit our task and train them on our dataset well.

Urban Layout Metrics

Urban layout metrics are utilized to measure whether the regenerated urban layouts match the specified layout description, considering the distinct characteristics of various road and building types. It is important to note that the regenerated outcomes need to undergo vectorization prior to the computation of urban layout metrics. The Road Layout Similarity (RLS) (AlHalawani et al. 2014) quantifies the extent to which the regenerated road layouts correspond to the attributes of the designated target road layouts. This is calculated as the ratio of the total count of regenerated roads minus the total count of actual target roads, divided by the number of testing samples, represented by:

$$RLS = \frac{\sum_t^N |\mathrm{num}\,(M_r) - \mathrm{num}\,(G_r)|}{N} \tag{1}$$

where M_r and G_r are the regenerated and GT target urban road layouts. $num(\cdot)$ is the total number of roads. N is

Figure 1: Examples of the 12 functional types.

Function		Region		Road	Building
		Total Area	*Count*	*Density*	*Count*
Residential	(RES)	608.34	12,285	Dense	604,082
Industrial	(IND)	151.02	1,789	Sparse	13,596
Commercial	(COMM)	105.18	2,694	Normal	21,309
Mixed-use	(MU)	98.25	22,388	Normal	46,280
Recreational	(REC)	97.75	33,416	Normal	120,541
Natural	(NAT)	72.60	9,473	Sparse	18,689
Retail	(RETAIL)	65.53	5,693	Normal	56,021
Public-related	(PUB)	44.39	41,920	Sparse	58,044
Education	(EDU)	22.83	1,199	Sparse	3,373
Healthcare	(HEALTHC)	18.74	2,124	Sparse	5,041
Station	(STN)	18.19	614	Sparse	1,868
Sustenance	(SUS)	14.54	13,829	Normal	29,197

Table 1: The statistics of our dataset. *Total Area* - the total area (km^2) occupied by each function. *Density* - the density level of the road layout belonging to each function. We rank the level based on (Zhang et al. 2015). *Count* - the total number of regions or buildings belonging to a function

.

the total number of testing samples. Building layout similarity (Chen et al. 2021) evaluates the morphological and spatial similarity of buildings between regenerated and ground-truth building layouts. We first model the building layout using the Delaunay triangulation (DT) graph (Lee and Lin 1986), where nodes are morphological properties (shapes) and edges are spatial relations. We then transform the DT graph from the spatial domain to the frequency domain based on Graph Fourier Transform. Finally, we compute BLS as the deviation between regenerated and ground-truth target building layouts w.r.t. the frequency domain features as:

$$\begin{aligned} \mu &= (\mu_{o1} - \mu_{g1})^2 + \cdots + (\mu_{on} - \mu_{gn})^2 \\ &\quad + (\mu_{om-n+1})^2 + \cdots + (\mu_{gm})^2 \\ BLS &= \sqrt{\mu}, \end{aligned} \tag{2}$$

where μ_o and μ_g are frequency signals for regenerated and GT target building layouts.

Diversity

Given the same layout description, target region and surrounding context, our method could regenerate diverse urban layouts as shown in Fig. 2.

References

AlHalawani, S.; Yang, Y.-L.; Wonka, P.; and Mitra, N. J. 2014. What Makes London Work like London? In *Proceedings of the Symposium on Geometry Processing*, SGP '14, 157–165. Goslar, DEU: Eurographics Association.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18208–18218.

Benes, B.; Zhou, X.; Chang, P.; and Cani, M.-P. R. 2021. Urban Brush: Intuitive and Controllable Urban Layout Editing. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, 796–814.

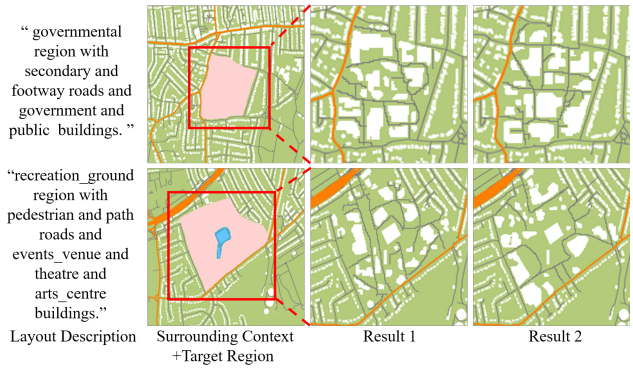

Figure 2: Diverse Results of Our Method. We regenerate diverse urban layouts of the target region (pink region) under the same layout description and surrounding context. We highlight the regenerated urban layouts of the red box.

Chen, G.; Esch, G.; Wonka, P.; Müller, P.; and Zhang, E. 2008. Interactive procedural street modeling. In *ACM SIGGRAPH*.

Chen, Z.; Ma, X.; Yu, W.; Wu, L.; and Xie, Z. 2021. Measuring the similarity of building patterns using Graph Fourier transform. *Earth Science Informatics*, 14: 1953–1971.

for Europe, U. N. E. C.; et al. 2010. Illustrated Glossary for Transport Statistics. Technical report, European Commission.

Lee, D.-T.; and Lin, A. K. 1986. Generalized Delaunay triangulation for planar graphs. *Discrete & Computational Geometry*, 1(3): 201–217.

OpenStreetMap contributors. 2017. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org.

Parish, Y.; and Müller, P. 2001. Procedural modeling of cities. In *ACM SIGGRAPH*, 301–308.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2085–2094.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Zhang, L.; Chen, Q.; Hu, B.; and Jiang, S. 2020. Text-guided neural image inpainting. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1302–1310.

Zhang, Y.; Li, X.; Wang, A.; Bao, T.; and Tian, S. 2015. Density and diversity of OpenStreetMap road networks in China. *J. Urban Manag.*, 4(2): 135–146.