Text2City: One-Stage Text-Driven Urban Layout Regeneration

Yiming Qin[1,2], **Nanxuan Zhao**[3*], **Bin Sheng**[1†], **Rynson W.H. Lau**[2*]

[1]Shanghai Jiao Tong University [2]City University of Hong Kong [3]Adobe Research
{yiming_qin, shengbin}@sjtu.edu.cn, nanxuanzhao@gmail.com, Rynson.Lau@cityu.edu.hk

Abstract

Regenerating urban layout is an essential process for urban regeneration. In this paper, we propose a new task called text-driven urban layout regeneration, which provides an intuitive input modal - text - for users to specify the regeneration, instead of designing complex rules. Given the target region to be regenerated, we propose a one-stage text-driven urban layout regeneration model, *Text2City*, to jointly and progressively regenerate the urban layout (*i.e.,* road and building layouts) based on textual layout descriptions and surrounding context (*i.e.,* urban layouts and functions of the surrounding regions). Text2City first extracts road and building attributes from the textual layout description to guide the regeneration. It includes a novel one-stage joint regenerator network based on the conditioned denoising diffusion probabilistic models (DDPMs) and prior knowledge exchange. To harmonize the regenerated layouts through joint optimization, we propose the interactive & enhanced guidance module for self-enhancement and prior knowledge exchange between road and building layouts during the regeneration. We also design a series of constraints from attribute-, geometry- and pixel-levels to ensure rational urban layout generation. To train our model, we build a large-scale dataset containing urban layouts and layout descriptions, covering 147K regions. Qualitative and quantitative evaluations show that our proposed method outperforms the baseline methods in regenerating desirable urban layouts that meet the textual descriptions.

Introduction

Urban regeneration is a crucial process to revitalize decaying or underused regions in the city (Amirtahmasebi et al. 2016), the basis of which is urban layout regeneration. In urban layout regeneration, plans are first constructed as white papers. Professional designers visualize the urban layout based on the white papers, and create the urban layout map by hand or with the aid of computer-assisted tools. Previous computer-assisted city modeling methods (Parish and Müller 2001; Chen et al. 2008; Groenewegen et al. 2009; Niese et al. 2022) generate urban layouts by employing complex hand-crafted rules. Users have to carefully adjust the control parameters, such as the number of roads, patterns, longest length, to obtain the desirable urban layout. These methods are time-consuming and not friendly for lay users. Thus, in this paper, we aim to allow the user to use textual descriptions to specify the layout requirements. Despite the great potential, text-driven urban layout regeneration poses several challenges. First, urban layout regeneration is complicated, involving many factors such as the diversity of layout characteristics, the richness of types (*e.g.,* function, road and building types), and the influence of the surrounding context. Second, roads and buildings are not independent to each other, but interact with each other during layout regeneration. For example, while roads should bypass historic buildings that need to be preserved, building footprints should generally align with existing roads. However, existing city modeling methods take a two-stage approach (Parish and Müller 2001; Chen et al. 2008; Groenewegen et al. 2009; Vanegas et al. 2010; Benes et al. 2021; Niese et al. 2022). They first generate roads and then rough building footprints according to the roads. They do not consider the interaction between road and building. Third, urban layout regeneration with text inputs requires additional annotations. How-

Figure 1: Text-driven Urban Layout Regeneration. Given the target region (pink region in (a)), Text2City regenerates its urban layout based on the textual layout description and the surrounding context. Orange and gray lines represent main and minor roads. White polygons represent buildings. Purple lines and blue polygons represent roads and buildings to be preserved, and are provided by the user as constraints. (b)-(d) are regenerated layouts of the red box in (a), based on different input textual layout descriptions.

*Nanxuan Zhao and Rynson Lau lead this project.
†Corresponding author.

ever, current urban layout datasets lack textual layout descriptions, *e.g.,* function, road type and building type.

To address these challenges, in this paper, we propose Text2City, a one-stage text-driven urban layout regeneration method. Given a target region to be regenerated, Text2City jointly and progressively regenerates road and building layouts of the target region conditioned on the textual layout descriptions and the surrounding context, as shown in Fig. 1. A textual layout description may include the target function, the road type, and the building type. While the target function specifies the land usage of the target region, the road and building types further provide fine-grained guidance. Text2City has two components: a Text-to-Attribute (T2A) network and a novel joint regenerator network. The T2A network is based on sentence-BERT (Reimers and Gurevych 2019) and Contrastive Language-Image Pretraining (CLIP) (Radford et al. 2021) to help decouple a layout description into road and building layout parts, and then extract layout attributes from them. The extracted layout attributes are then utilized to guide our joint regenerator to regenerate the urban layout to match with the input textual layout description.

Our one-stage joint regenerator network is designed to regenerate the target urban layout based on conditioned DDPMs (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021; Avrahami, Lischinski, and Fried 2022) and prior knowledge exchange. The regenerator comprises two streams, the *building-assisted road layout (BaRL) regeneration stream* and the *road-assisted building layout (RaBL) regeneration stream*. Joint optimization is achieved through our interactive & enhanced guidance (IEG) module. In the BaRL regeneration stream, the IEG module performs self-enhancement and employs prior knowledge of building layouts to harmonize the regenerated road layout with the existing building layout. Similarly, in the RaBL regeneration stream, the IEG module also performs self-enhancement and employs prior knowledge of road layouts to assist the building layout regeneration. Besides, the regeneration of an urban layout is also affected by its surrounding context, especially near the boundary. Hence, we progressively sample and apply the surrounding context to the joint regeneration process to provide local details. We also design a set of constraints to optimize regenerated urban layouts from attribute-, geometry-, and pixel-levels to obtain rational results.

Finally, we have collected a large-scale urban layout dataset covering 147K regions with rich textual annotations including functions, road layouts and building layouts. We evaluate our proposed method and baselines for text-driven urban layout regeneration on our dataset, and the results show that our proposed method outperforms all baselines.

To summarize, our main contributions are:

- We propose Text2City for the new task: text-driven urban layout regeneration. Our proposed method is able to regenerate rational urban layouts from textual descriptions and surrounding contexts.
- We propose a novel one-stage joint regenerator network based on the conditioned DDPMs and prior knowledge exchange, where the IEG module utilizes the prior knowledge of road and building layouts for self-enhancement and joint optimization. We also design a set of constraints at attribute-, geometry- and pixel-levels to ensure valid urban layout synthesis.
- We have collected a large-scale urban layout dataset covering 147K regions with rich annotations, including functions, road/building layouts, and text descriptions.
- We demonstrate the effectiveness of our method on text-driven urban layout regeneration through extensive experiments. We also showcase applications of our method with other user constraints and in other cities.

Related Work

Urban Layout Design. Previous works mainly consider city modeling using procedural modeling methods. These methods (Parish and Müller 2001; Chen et al. 2008; Groenewegen et al. 2009; Weber et al. 2009; Lipp et al. 2011) required expertise to design the parameters to obtain the desired urban layout manually. UrbanBrush (Benes et al. 2021) provided a layout editing tool for users to set the brush parameters, *e.g.,* impact region, population, height and amount.

In recent years, deep learning-based methods are introduced to urban road layout design. StreetGAN (Hartmann et al. 2017) is an example-based road layout generation method. Chu *et al.* (Chu et al. 2019) treated a road layout as a graph. Starting from scratch, they utilized a sequential generative model to generate the road network iteratively.

However, traditional methods generate urban layouts in a two-stage manner which do not consider the interaction between roads and buildings, and always start from scratch ignoring the surrounding context. In addition, designing complex rules for such methods requires professional knowledge, making it difficult for novices. Latest deep-learning based methods focus on generating style-based urban road layouts, such as London style or New York style. Unlike previous works, text-driven urban layout regeneration aims to regenerate the urban layout of a target region surrounded by some existing regions and conditioned on the target layout description. To address this task, we introduce a one-stage text-driven deep-learning-based generative method that learns the complex regeneration process from the layout descriptions and the surrounding context.

Text-driven Methods. Some works (Zhang et al. 2016; Nichol et al. 2021; Ramesh et al. 2021; Saharia et al. 2022; Lugmayr et al. 2022) explored text-driven image generation. They utilized text as conditions to generate target images. Some other works (Nam, Kim, and Kim 2018; Liu et al. 2020; Patashnik et al. 2021; Bau et al. 2022; Avrahami, Lischinski, and Fried 2022) studied text-driven image manipulation. They modified images globally or locally conditioned on the text descriptions. CLIPdraw (Frans, Soros, and Witkowski 2021) generated drawings by maximizing the similarity between the given description and the generated drawing based on the CLIP model. Jiang *et al.* (Jiang et al. 2022) proposed a text-driven controllable human image generation method. Given a human pose and text, they first translated the text into one-shot attributes, and then obtained the human parsing map and the human image with clothing shapes and textures. Michel *et al.* (Michel et al. 2022) utilized CLIP and the neural style filed network to stylize

a 3D mesh based on a target text prompt. Text2LIVE (Bar-Tal et al. 2022) could edit the appearance of existing objects or augment the scene with visual effects of an image or a video based on the text prompt. Jain *et al.* (Jain et al. 2022) proposed a text-driven method to synthesize diverse 3D objects. Most of the previous works utilized CLIP to achieve the text-driven task.

Our work also utilizes the CLIP model to extract layout attributes from the given textual layout description to guide our proposed joint regenerator for urban layout regeneration.

Our Dataset

The Need for a Dataset. In our task, we use the CLIP model to extract layout attributes and guide the urban layout regeneration process. However, an experiment conducted by us on ClipCap (Mokady, Hertz, and Bermano 2021) shows that pre-trained CLIP has poor generalization to urban layouts. Unfortunately, current publicly available datasets (Services 2016; Belli and Kipf 2019) do not have detailed urban layout annotations. These problems inspire us to create a large-scale urban layout dataset with rich annotations. In summary, our dataset has the following properties. (1) It contains 147K regions covering most areas of the Greater London in the UK. (2) We annotate region, road and building layout information. (3) Each region is annotated with a textual description of the urban layout.

Data Collection. We collect our data from Open Street Map (OSM) (OpenStreetMap contributors 2017), which contains diverse urban map data contributed by people worldwide. The Greater London area is chosen in this work. The raw data are in geographic format, containing sequences of (latitude, longitude) coordinates and textual annotations.

Data Processing. We first remove noisy data by merging adjacent regions with the same function (*i.e.,* land use) and eliminating overlapping roads and buildings. We then render layout data from geographic format to layered image format, using the style of urban layout images on the web (Radford et al. 2021) for inspiration. The region layer is rendered with region shape and function type. The road layer is rendered based on the road hierarchy (for Europe et al. 2010), with different types of road having different widths and colors. The building layer is rendered from an aerial view based on the shape. A textual layout description is then created for each region, combining the function type, road type and building type that this region occupies. Finally, we obtain a dataset of (layout map, text) pairs for text-driven urban layout regeneration. Please see supplementary for details.

Text2City

Our objective is to regenerate urban layouts conditioned on textual layout descriptions and the surrounding context in one stage. Given a target region R, the target urban layout description T and the surrounding context $S = (S_r, S_b)$, Text2City outputs the regenerated urban layout M, containing the road M_r and building M_b layout. Note that the output regenerated urban layout M combines the regenerated urban layout in the target region and the existing surrounding layout. The whole pipeline of Text2City is illustrated in Fig. 2. We start by extracting urban layout attributes from the textual descriptions T. To account for the interaction between roads and buildings during regeneration, we present the one-stage joint regenerator. Additionally, we present the constraints used to enhance urban layout regeneration.

The Text-to-Attribute (T2A) Network. As mentioned above, the textual layout description consists of the target function, road type and building type, where the target function determines the global characteristics of the target layout, and the road and building types provide fine-grained guidance. For example, "residential region with pedestrian roads and residential and house buildings": "residential region" always has narrow, short roads and crowded buildings. The "pedestrian roads" indicates the requirement for narrow roads, while "residential and house buildings" suggests the need for both crowded ordinary and sparse high-class residences. Note that the target function is mandatory but the road and building types are optional. Since the road and building layouts have distinct properties and structures (road layouts comprise lines with varying widths and colors while building layouts comprise polygons with diverse shapes), we first decouple the layout description into road and building layouts. We use the sentence-BERT (Reimers and Gurevych 2019) to identify keywords related to "region", "road", and "building" to split the description into road layout (T_r) and building layout (T_b) parts. T_r and T_b both contain the target function of the region because the target function determines the global characteristics. Subsequently, T_r and T_b are fed into the text encoder of a fine-tuned CLIP model and obtain the attributes for road layout (a_r) and building layout (a_b) separately. The extracted attributes are used to guide our one-stage joint regenerator.

The Joint Regenerator Network. We aim to jointly and progressively regenerate the road layout and building layout conditioned on the extracted attributes and the surrounding context. To achieve this, we propose a one-stage joint generative model based on the conditioned DDPMs and prior knowledge exchange, as shown in Fig. 2. In the forward noising process, we introduce Gaussian noise with variance $\beta^t \in (0, 1)$ to the urban layout map $M^0 = M$ to obtain a series of noisy urban layout map M^t with t steps as follows,

$$\begin{aligned} q\left(M^1, \ldots, M^K \mid M^0\right) &= \prod_{t=1}^{K} q\left(M^t \mid M^{t-1}\right), \\ q\left(M^t \mid M^{t-1}\right) &= \mathcal{N}\left(M^t; \sqrt{1-\beta^t} M^{t-1}, \beta^t \mathbf{I}\right). \end{aligned} \tag{1}$$

If K is a large number, M^K is a nearly standard Gaussian noise $\mathcal{N}(0, \mathbf{I})$. Further, we can also directly sample M^t from M^0 without the intermediate steps,

$$\begin{aligned} q(M^t \mid M^0) &= \mathcal{N}(M^t; \sqrt{\bar{\alpha}^t} M^0, (1-\bar{\alpha}^t)\mathbf{I})), \\ M^t &= \sqrt{\bar{\alpha}^t} M^0 + \sqrt{(1-\bar{\alpha}^t)}\epsilon, \end{aligned} \tag{2}$$

where $\bar{\alpha}^t = \prod_{s=0}^{t} \alpha^s$ is the total noise variance, $\alpha^t = 1 - \beta^t$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Based on these, the regeneration process is the reverse of the above. We denoise the noisy version M^t using the learned deep neural network p_θ as follows,

$$p_\theta(M^{t-1} \mid M^t) = \mathcal{N}(M^{t-1}; \mu_\theta(M^t, t), \Sigma_\theta(M^t, t)), \tag{3}$$

where $\mu_\theta(\cdot)$ and $\Sigma_\theta(\cdot)$ are the parameters of the predicted Gaussian distribution from the p_θ. The regenerated urban

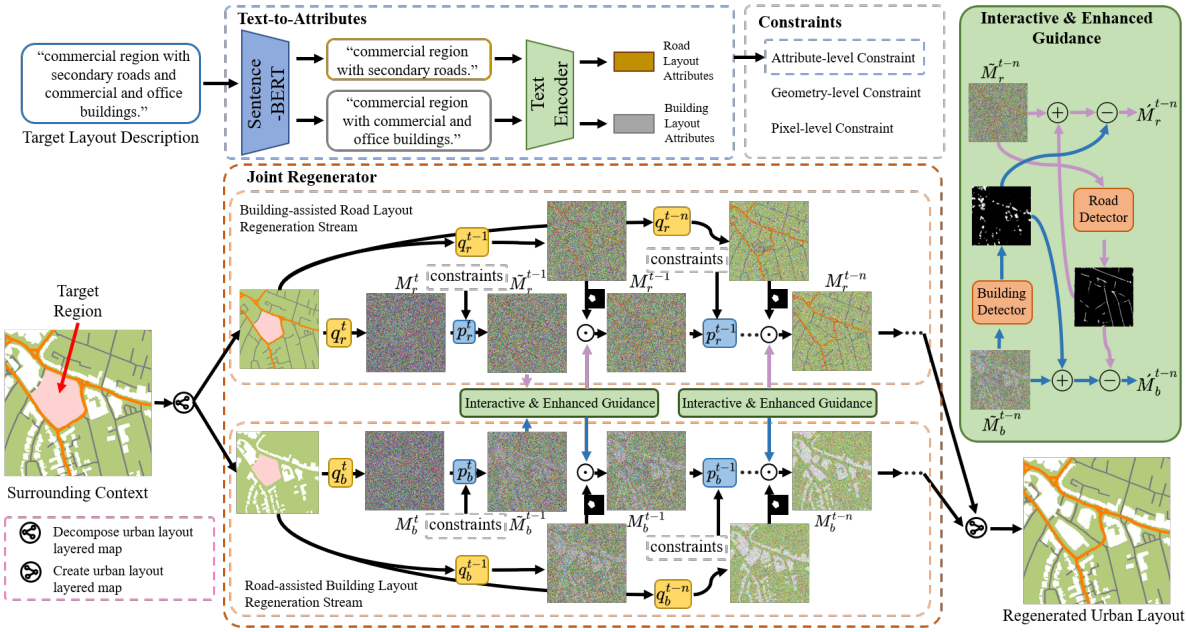

Figure 2: Overview of Text2City. Given the target region, target layout description and the surrounding context, Text2City outputs the regenerated urban layout. First, the text-to-attribute (T2A) network decouples and extracts the layout attributes to guide the regeneration. Then, our joint regenerator network based on conditioned DDPMs and prior knowledge exchange jointly and progressively regenerates the urban road and building layouts, while being conditioned by the attributes and the surrounding context. The interactive & enhanced guidance (IEG) module facilitates prior knowledge exchange between road and building layouts, allowing for self-enhancement and joint optimization. In addition, a set of constraints are used to ensure the urban layout is desirable and rational.

layout $\hat{M}^0$ can be obtained from Eq. 2 as follows,

$$\hat{M}^0 = \frac{M^t}{\sqrt{\bar{\alpha}^t}} - \frac{\sqrt{1-\bar{\alpha}^t}\epsilon_\theta\left(M^t, t\right)}{\sqrt{\bar{\alpha}^t}}, \tag{4}$$

where $\epsilon_\theta(\cdot)$ is the predicted noise from the network. In BaRL stream, we first obtain the regenerated road layout $\tilde{M}_r^{t-1}$ through DDPM conditioned on the target road attributes a_r as follows,

$$\tilde{M}_r^{t-1} \sim \mathcal{N}(\tilde{M}_r^{t-1}; \mu_\theta(M_r^t, t) + \Sigma_\theta(M_r^t, t)\nabla_r, \Sigma_\theta(M_r^t, t)), \tag{5}$$

where ∇_r is the gradient calculated by the target road layout attributes a_r and the regenerated road layout as follows,

$$\nabla_r = \frac{1}{N}\sum_{i=1}^{N} \nabla_{\hat{M}_r^{0,\text{aug}}} \mathcal{L}_a(\hat{M}_r^{0,aug}, a_r, R), \tag{6}$$

where $\mathcal{L}_a$ is Eq. 9. $\hat{M}_r^{0,aug}$ are extending augmentation samples through (Avrahami, Lischinski, and Fried 2022) to mitigate the problems that the adversarial noise damages the performance of the CLIP model. The extending augmentation method extends $\hat{M}_r^0$ to N samples through projection transformation. Therefore, the final gradient is obtained from the average gradients of N samples instead of a single.

To get rational layouts, we propose the IEG module for self-enhancement and joint optimization between road and building layouts to bridge two streams. The IEG module first obtains candidate heatmaps for roads and buildings using the road and building detectors based on current regenerated urban layouts. The road detector $\mathcal{D}_R$ based on (Pautrat et al. 2021) obtains the road heatmap and the building detector $\mathcal{D}_B$ inspired by the model (Yang et al. 2016) generates the building heatmap, as shown in Fig. 2. In the BaRL regeneration stream, IEG module leverages the road heatmap H_r^{t-1} combined with the sampled road r^{t-1} for self-enhancement. The prior knowledge of building layout dictates that regenerated roads cannot cross buildings but lead to them. This understanding informs the joint optimization process, aligning the regenerated road layout with the regenerated building layout as follows:

$$\acute{M}_r^{t-1} = \tilde{M}_r^{t-1} + r^{t-1} \odot H_r^{t-1} - H_b^{t-1} \odot r^{t-1}. \tag{7}$$

Thus, we obtain the optimized road layout $\acute{M}_r^{t-1}$ harmonized with the existing building layout. Furthermore, the regenerated urban layout is also influenced by the surrounding context. We sample the surrounding context and apply it to the stream to provide the local details as follows,

$$\begin{aligned} M_r^{t-1} &= \acute{M}_r^{t-1} \odot R + S_r^{t-1} \odot (1-R), \\ S_r^{t-1} &\sim \mathcal{N}(S_r^{t-1}; \sqrt{\bar{\alpha}^{t-1}} S_r^0, (1-\bar{\alpha}^{t-1})\mathbf{I})). \end{aligned} \tag{8}$$

Note that blending the $\tilde{M}_r^{t-1}$, r^{t-1} and S_r^{t-1} produces a result outside the current manifold, which is corrected in the next diffusion step that projects the result to the $t-2$ step manifold, thereby ensuring coherence. Similarly, in RaBL regeneration stream, we utilize the prior knowledge of the regenerated road layout and the surrounding context S_b^{t-1} to optimize the regenerated building layout M_b^{t-1}. The prior knowledge of road layout is that the regenerated buildings should be along the roads. By iteratively repeating the above joint regeneration process, we obtain the regenerated urban layout M including the regenerated road layout $M_r = M_r^0$ and building layout $M_b = M_b^0$. Additionally, the urban layout could be vectorized like (Chu et al. 2019), extending its

representation beyond the current resolution.

Constraints. To ensure urban layouts with desirable characteristics and rational geometric structures, we design a set of constraints on three domains: attribute, geometry and pixel. ***Attribute-level:*** The attribute-level constraint aims to align the regenerated urban layout in the target region with the intended layout description. We compute the cosine distance between the attributes and the regenerated urban layout embedding and get the constraint as follows,

$$\mathcal{L}_{\mathrm{a}} = 1 - \frac{aCLIP_{img}(M \odot R)}{|a||CLIP_{img}(M \odot R)|}, \tag{9}$$

where $CLIP_{img}$ represents the image encoder of CLIP. M denotes the road or building layout, a is the road or building attributes in different streams. ***Geometry-level:*** The Isoperimetric Quotient (IQ) gauges the compactness of a shape, where a higher value indicates greater compactness. During building layout regeneration, we leverage the IQ to enhance the building shape, formulated as follows,

$$\mathcal{L}_g = -\frac{4\pi A(\mathcal{D}_B(M_b^t))}{P(\mathcal{D}_B(M_b^t))^2}, \tag{10}$$

where $P(\cdot)$ is the perimeter of the building footprint, and $A(\cdot)$ is the area covered by the building. $\mathcal{D}_B$ is our building detector to get the building footprints at step t. This constraint punishes irregular and fragmented building shapes in the building layout regeneration. ***Pixel-level:*** We use the $\mathcal{L}_2$ loss as the range loss to control how far out-of-range pixel values are allowed to be as follows,

$$\mathcal{L}_r = \|M^t - Clamp(M^t)\|_2, \tag{11}$$

where $Clamp(\cdot)$ is used to clamp pixel values to $[-1, 1]$. This constraint encourages the regenerator to predict the proper pixel values.

Above all, the total loss is set to the weighted sum of attribute-, geometry- and pixel-level constraints as follows,

$$\mathcal{L}_{all} = \lambda_a \mathcal{L}_a + \lambda_g \mathcal{L}_g + \lambda_r \mathcal{L}_r, \tag{12}$$

where λ_a, λ_g and λ_r are the weight coefficients. In our experiment, we set $\lambda_a = 100$, $\lambda_g = 70$ and $\lambda_r = 5$.

Experiment and Evaluation

In this section, we compare our method with state-of-the-art (SOTA) methods both qualitatively and quantitatively. Besides, a user study is conducted to evaluate the performance of our method. Furthermore, we demonstrate the effectiveness of our method through an ablation study.

Implementation Details. Our method is implemented using Pytorch. We fine-tune the pre-trained CLIP model on our dataset with a batch size of 64 and an initial learning rate of $1e-8$. The Adam optimizer is employed with β_1 set to 0.9 and β_2 set to 0.98. For our experiments, each urban layout map covers $1km^2$ and the central region with a single function is taken as the target region. On an RTX 2080ti, it takes about one and a half minutes to complete a text-driven urban layout regeneration using our method.

Qualitative and Quantitative Comparison. To the best of our knowledge, our work is the first to handle text-driven urban layout regeneration. We first compare our method with SOTA traditional city modeling methods: CityEngine (Parish and Müller 2001), IPSM (Chen et al. 2008) and UrbanBrush (Benes et al. 2021). All these methods require manual adjustments with expertise. We first generate the road network and then the region is divided into rough building footprints based on the road network. Finally, we manually merge the generated urban layout with the surrounding real urban layout. As we render the urban layout in a bird-view image, we also compare our method with SOTA text-driven image synthesis methods: TDANet (Zhang et al. 2020), BD (Avrahami, Lischinski, and Fried 2022), StyleCLIP (Patashnik et al. 2021) and SDv2 (Rombach et al. 2022). We modify these methods to suit our task; see supplementary for details.

Qualitative results are shown in Fig. 3. Traditional methods possess limitations in generating suitable urban layouts due to intricate hand-drafted parameters. Incoherence manifests at the boundary due to manual merging operations and the absence of the surrounding context as a condition. Moreover, these methods can not generate detailed building shapes. Both TDANet and StyleGAN fail to regenerate the urban layouts. BD and SDv2 can regenerate urban layouts but struggle to regenerate rational geometric structures of urban layouts. Moreover, they face difficulties in regenerating urban layouts with sparse surrounding contexts, as seen in the second row of Fig. 3. Among them, our method achieves the best results, regenerating desired and rational urban layouts. The main reason for the failure of baselines is that they are designed for natural images, which have rich textual and structural information. In contrast, urban layouts have little textural and sparse structural information. Besides, without decoupling the complex layout description, it is difficult for these models to control the road and building characteristics well. Our method also regenerates diverse results, shown in Fig. 5 and supplementary.

We also evaluate the results quantitatively using generative and urban layout metrics. Generative metrics encompass SSIM (Wang et al. 2004), FID (Heusel et al. 2017) and Wasserstein Distance (WD) (Arjovsky, Chintala, and Bottou 2017) to measure the quality and diversity. Urban layout metrics measure whether the regenerated urban layouts match the specified layout description, considering the distinct characteristics of various road and building types in the layout description. Road layout similarity (RLS) measures how well the regenerated road layouts match the characteristics of the target (AlHalawani et al. 2014). Building layout similarity (BLS) evaluates the morphological and spatial similarity between regenerated and target buildings (Chen et al. 2021). See supplementary for details. Note that the regenerated urban layouts need to be vectorized before computing urban layout metrics. The results, shown in Tab. 1, indicate that our method outperforms all baselines on all metrics, particularly in urban layout metrics.

User study. We also conduct a user study to evaluate our method and baselines (TDANet, BD and SDv2). Since TDANet and StyleCLIP both fail to regenerate urban layouts, we choose one for the user study. We recruit 20 participants and design 15 questions to assess the rationality of the regenerated urban layout (Rationality), harmony with the surrounding layout (Harmony) and match with the layout description (Match). Each question contains four examples generated by baselines and our method, and the partic-

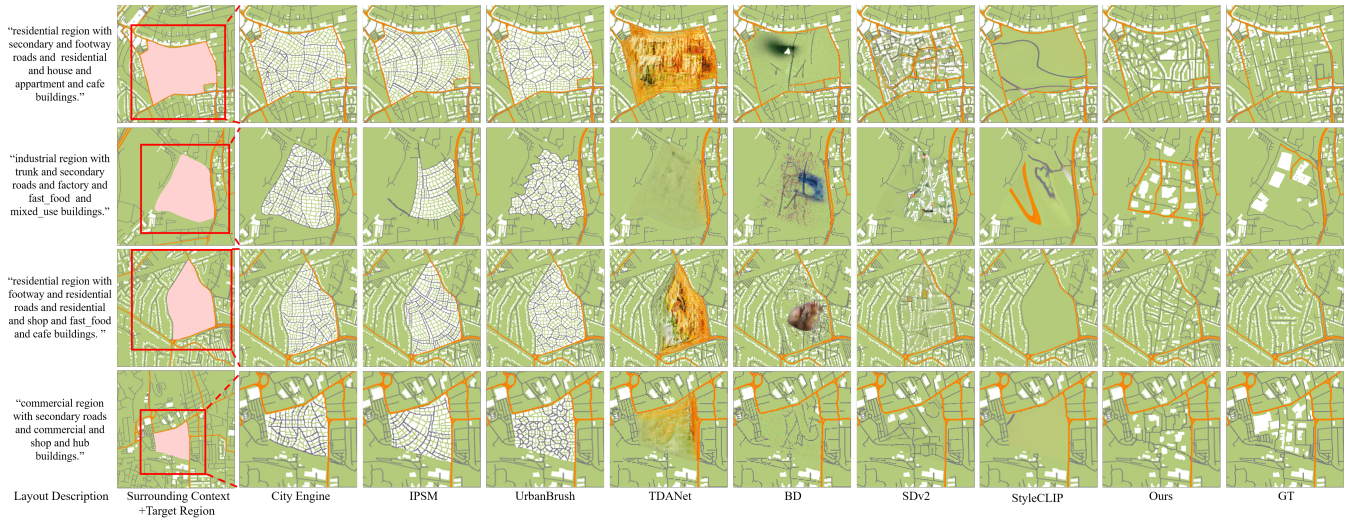

Figure 3: Qualitative Comparison. Our method achieves better performance than baselines on text-driven urban layout regeneration. The pink region is the target region. We highlight the regenerated urban layouts of the red box.

Method	SSIM ↑	FID ↓	WD ↓	RLS ↓	BLS ↓
CityEngine	0.51	145.54	0.20	262.80	0.93
IPSM	0.54	137.11	0.18	241.70	0.87
UrbanBrush	0.51	147.73	0.17	317.40	0.92
TDANet	0.40	161.15	0.20	-	-
StyleCLIP	0.31	170.71	0.26	-	-
BD	0.64	130.54	0.14	110.60	0.98
SDv2	0.72	120.62	0.10	56.19	0.60
Ours	**0.89**	**77.71**	**0.04**	**19.37**	**0.24**

Table 1: Quantitative Results. We compute SSIM (higher is better), FID (lower is better), and WD (lower is better), RLS (lower is better), BLS (lower is better). Our method outperforms all baselines on all metrics. − means it fails to generate urban layouts.

Method	Rationality↑	Harmony ↑	Match ↑
TDANet	1.36 (0.17)	1.65 (0.20)	1.46 (0.14)
BD	2.46 (0.35)	2.21 (0.25)	2.16 (0.28)
SDv2	2.67 (0.17)	2.46 (0.10)	2.77 (0.09)
Ours	**3.38** (0.13)	**3.08** (0.09)	**3.61** (0.05)

Table 2: User Study Results. Our method achieved the highest score. The variance of the score is inside the braces.

ipants rank examples in descending order. The final score is the ranking-weighted score, which is calculated by ranking weight and count. The ranking order determines the ranking weight. For example, ranking weight 4 is assigned to the first rank and 2 to the third rank. The count is the number of times the method occurs at a specific ranking order. The higher scores indicate better results. As shown in Tab. 2, our method receives the highest score.

Ablation Study. We first validate the necessity of our dataset by performing the experiment with the original pre-trained CLIP, *i.e.,* ours w/o fine-tune. Results in Fig. 4(c) indicate that the original pre-trained CLIP has poor generalization on the urban layout and layout description. Hence, it is necessary to have a dataset for our task. Next, we evaluate the T2A network and regenerate urban layout directly conditioned on the original layout description, *i.e.,* ours w/o T2A. As shown in Fig. 4(d), the coupled attributes lead to inadequate guidance. We then evaluate the IEG module by removing it, *i.e.,* ours w/o IEG. Thus, we regenerate the road layout and building layout in two streams without joint optimization. Results in Fig. 4(e) show that broken roads and overlapping layouts appear. Besides, buildings appear incorrect shapes and distributions without prior knowledge of roads (see red arrows in Fig. 4). We then evaluate our geometry-level constraint and remove $\mathcal{L}_g$, *i.e.,* ours w/o $\mathcal{L}_g$. Results in Fig. 4(f) show irregular and broken building shapes. We also remove the road self-enhancement, *i.e.,* ours w/o SR. Results in Fig. 4(g) indicate that the road layout may be broken or missing, especially in the center of the large target region. Quantitative results in Tab. 3 further prove the effectiveness of our proposed components.

Method	SSIM ↑	FID ↓	WD ↓	RLS ↓	BLS ↓
Ours w/o fine-tune	0.65	126.72	0.13	130.44	0.98
Ours w/o T2A	0.68	122.42	0.11	152.22	0.89
Ours w/o IEG	0.76	115.11	0.08	58.39	0.44
Ours w/o SR	0.78	112.41	0.08	50.56	0.38
Ours w/o $\mathcal{L}_g$	0.81	106.68	0.07	35.17	0.61
Ours	**0.89**	**77.71**	**0.04**	**19.37**	**0.24**

Table 3: Quantitative Results of Ablation Study. The results show that the proposed components benefit our task.

Applications

This section describes some applications of our method in urban layout regeneration.

User Constraints. Our method allows users to add additional constraints. ***Main Road-Guided Urban Layout Regeneration:*** Urban planners commonly undertake urban layout regeneration with a foundation in the pre-existing main roads (Amen and Nia 2020; Zhang, Zhang, and Yin 2021).

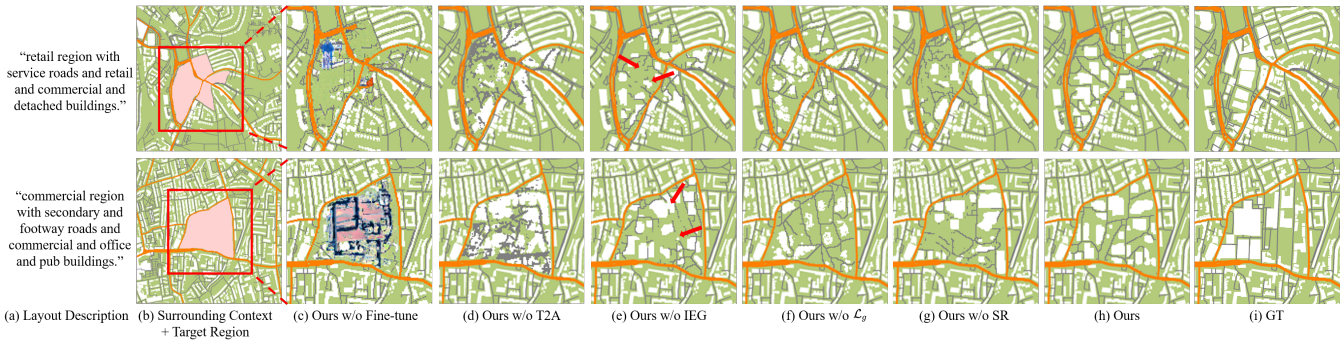

Figure 4: Ablation Study. We remove the components of our model and regenerate the urban layout again. All of our proposed components have positive contributions to our method. We highlight the regenerated results of the red box.

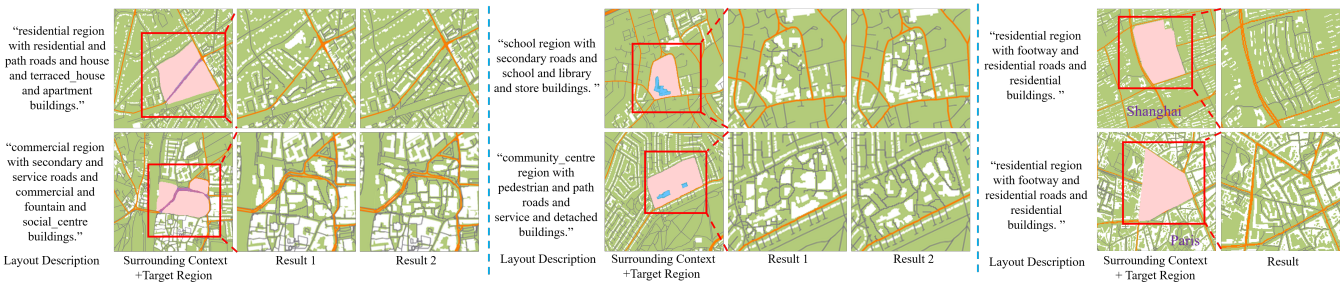

Figure 5: Applications. Left: main road-guided urban layout regeneration. Middle: building-preserved urban layout regeneration. Right: Other cities. The pink regions are the target regions, the purple lines are the unchanged main roads and the blue polygons are the preserved buildings. We highlight the regenerated results of the red box.

Our method allows users to add the pre-existing main roads as constraints like the surrounding context. Eventually, our method jointly regenerates the urban layouts that match the unchanged main roads as shown in Fig. 5 left.

Building-preserved Urban Layout Regeneration: It is also important to preserve significant buildings such as historical sites and public buildings (Parlewar and Fukukawa 2006). Users can introduce preserved buildings as constraints within our method. The context of preserved buildings is gradually integrated into the layout regeneration under prior knowledge exchange. The results are shown in Fig. 5 middle.

Other Cities. We collect Paris and Shanghai data from OSM and fine-tune our model on these data. Fig. 5 left indicates that our method can regenerate urban layouts for other cities with city style, which shows the generalization of our model.

Limitations and Future Work

One limitation is the long inference time due to denoising and joint optimization calculation. Thus, research in accelerating sampling is needed. If there are uncommon layout descriptions, CLIP may not guide the joint regenerator effectively, resulting in suboptimal urban layouts. Besides, due to the dataset sourced from OSM, we cannot specify low-level controls like road length/style and building number. To overcome these challenges, expanding the dataset to cover a wider range of urban layouts and more annotations is essential.

We intend to extend text-driven urban layout regeneration to 3D urban layouts for further research. Besides, fine-grained urban layout regeneration, such as the placement of traffic signs and street lights, is another avenue for research.

Conclusion

In this paper, we focus on a new task, text-driven urban layout regeneration. We propose Text2City to jointly and progressively regenerate the urban layout conditioned on the target layout description and surrounding context. We first extract road and building layout attributes to guide joint regeneration. We then propose a novel one-stage joint regenerator based on conditioned DDPMs and prior knowledge exchange, where the IEG module utilizes prior knowledge of road and building layouts for self-enhancement and joint optimization. We also constrain the regeneration in attribute-, geometry- and pixel-levels. We have collected a large-scale dataset that contains 147K regions with rich annotations for our task. Experimental results show that our method outperforms all baselines in text-driven urban layout regeneration. The code and dataset will be available at https://github.com/LittleQBerry/Text2City.

Acknowledgements

This project is in part supported by a GRF grant from the Research Grants Council of Hong Kong (No.: 11205620), and the National Natural Science Foundation of China under grant number 62272298 and 62077037.

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.

Amen, M. A.; and Nia, H. A. 2020. The Effect of Centrality Values in Urban Gentrification Development: A Case Study of Erbil City. *Civil Engineering and Architecture*, 8(5): 916–928.

Amirtahmasebi, R.; Orloff, M.; Wahba, S.; and Altman, A. 2016. *Regenerating Urban Land: A Practitioner's Guide to Leveraging Private Investment*. Urban Development. World Bank Publications. ISBN 9781464804748.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223. PMLR.

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.

Bar-Tal, O.; Ofri-Amar, D.; Fridman, R.; Kasten, Y.; and Dekel, T. 2022. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, 707–723. Springer.

Bau, D.; Andonian, A.; Cui, A.; Park, Y.; Jahanian, A.; Oliva, A.; and Torralba, A. 2022. Paint by Word.

Belli, D.; and Kipf, T. 2019. Image-Conditioned Graph Generation for Road Network Extraction. In *NeurIPS Workshop on Graph Representation Learning*.

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.

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.

Chu, H.; Li, D.; Acuna, D.; Kar, A.; Shugrina, M.; Wei, X.; Liu, M.-Y.; Torralba, A.; and Fidler, S. 2019. Neural turtle graphics for modeling city road layouts. In *ICCV*, 4522–4530.

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

Frans, K.; Soros, L. B.; and Witkowski, O. 2021. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*.

Groenewegen, S. A.; Smelik, R. M.; de Kraker, K. J.; and Bidarra, R. 2009. Procedural city layout generation based on urban land use models. *Eurographics Short Paper Proceedings*.

Hartmann, S.; Weinmann, M.; Wessel, R.; and Klein, R. 2017. StreetGAN: towards road network synthesis with generative adversarial networks. *Václav Skala-UNION Agency*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, volume 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *CoRR*, abs/2006.11239.

Jain, A.; Mildenhall, B.; Barron, J. T.; Abbeel, P.; and Poole, B. 2022. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 867–876.

Jiang, Y.; Yang, S.; Qju, H.; Wu, W.; Loy, C. C.; and Liu, Z. 2022. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4): 1–11.

Lipp, M.; Scherzer, D.; Wonka, P.; and Wimmer, M. 2011. Interactive modeling of city layouts using layers of procedural content. *Computer Graphics Forum*, 30(2): 345–354.

Liu, Y.; Nadai, M. D.; Cai, D.; Li, H.; Alameda-Pineda, X.; Sebe, N.; and Lepri, B. 2020. Describe What to Change: A Text-guided Unsupervised Image-to-Image Translation Approach. *acm multimedia*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Michel, O.; Bar-On, R.; Liu, R.; Benaim, S.; and Hanocka, R. 2022. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13492–13502.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nam, S.; Kim, Y.; and Kim, S. J. 2018. Text-Adaptive Generative Adversarial Networks: Manipulating Images with Natural Language. *neural information processing systems*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Niese, T.; Pirk, S.; Albrecht, M.; Benes, B.; and Deussen, O. 2022. Procedural Urban Forestry. *ACM Transactions on Graphics (TOG)*, 41(2): 1–18.

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.

Parlewar, P.; and Fukukawa, Y. 2006. Urban regeneration of historic towns: regeneration strategies for Pauni, India. *The Sustainability City IV: Urban Regeneration and Sustainability*, 209–228.

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.

Pautrat, R.; Lin, J.-T.; Larsson, V.; Oswald, M. R.; and Pollefeys, M. 2021. SOLD2: Self-supervised occlusion-aware line description and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11368–11378.

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.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. *international conference on machine learning*.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487*.

Services, A. W. 2016. SpaceNet dump retrieved from https://registry.opendata.aws/spacenet.

Vanegas, C. A.; Aliaga, D. G.; Wonka, P.; Müller, P.; Waddell, P.; and Watson, B. 2010. Modelling the appearance and behaviour of urban spaces. In *Computer Graphics Forum*, volume 29, 25–42. Wiley Online Library.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Weber, B.; Müller, P.; Wonka, P.; and Gross, M. 2009. Interactive geometric simulation of 4d cities. *Computer Graphics Forum*, 28(2): 481–492.

Yang, J.; Price, B.; Cohen, S.; Lee, H.; and Yang, M.-H. 2016. Object contour detection with a fully convolutional encoder-decoder network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 193–202.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2016. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. *international conference on computer vision*.

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, L.; Zhang, R.; and Yin, B. 2021. The impact of the built-up environment of streets on pedestrian activities in the historical area. *Alexandria Engineering Journal*, 60(1): 285–300.