

CityDreamer: Compositional Generative Model of Unbounded 3D Cities

Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, Ziwei Liu 
 S-Lab, Nanyang Technological University

{haozhe.xie, zhaoxi001, fangzhou001, ziwei.liu}@ntu.edu.sg

<https://haozhexie.com/project/city-dreamer>

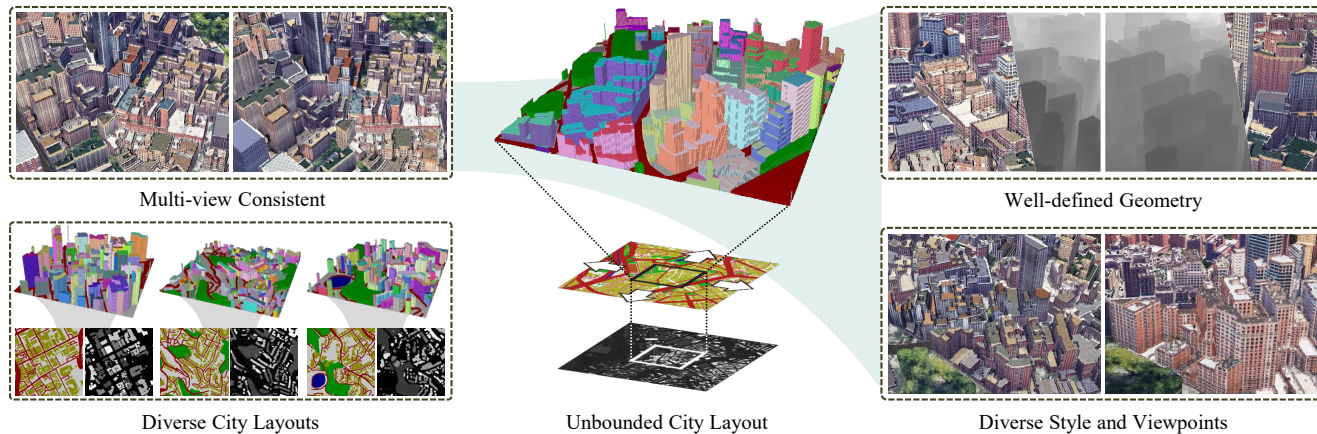


Figure 1. The proposed CityDreamer generates a wide variety of unbounded city layouts and multi-view consistent appearances, featuring well-defined geometries and diverse styles.

Abstract

3D city generation is a desirable yet challenging task, since humans are more sensitive to structural distortions in urban environments. Additionally, generating 3D cities is more complex than 3D natural scenes since buildings, as objects of the same class, exhibit a wider range of appearances compared to the relatively consistent appearance of objects like trees in natural scenes. To address these challenges, we propose **CityDreamer**, a compositional generative model designed specifically for unbounded 3D cities. Our key insight is that 3D city generation should be a composition of different types of neural fields: 1) various building instances, and 2) background stuff, such as roads and green lands. Specifically, we adopt the bird’s eye view scene representation and employ a volumetric render for both **instance-oriented** and **stuff-oriented** neural fields. The generative hash grid and periodic positional embedding are tailored as scene parameterization to suit the distinct characteristics of building instances and background stuff. Furthermore, we contribute a suite of **CityGen Datasets**, including OSM and GoogleEarth, which comprises a vast amount of real-world city imagery to enhance the realism of the generated 3D cities both in their layouts and appearances. CityDreamer achieves state-of-the-art performance not only in generating realistic 3D cities but also in localized editing within the generated cities.

1. Introduction

In the wave of the metaverse, 3D asset generation has drawn considerable interest. Significant advancements have been achieved in generating 3D objects [45, 57, 58], 3D avatars [24, 29, 60], and 3D scenes [7, 11, 34]. Cities, being one of the most crucial 3D assets, have found widespread use in various applications, including urban planning, environmental simulations, and game asset creation. Therefore, the quest to make 3D city development accessible to a broader audience encompassing artists, researchers, and players, becomes a significant and impactful challenge.

In recent years, notable advancements have been made in the field of 3D scene generation. GANCraft [22] and SceneDreamer [11] use volumetric neural rendering to produce images within the 3D scene, using 3D coordinates and corresponding semantic labels. Both methods show promising results in generating 3D natural scenes by leveraging pseudo-ground-truth images generated by SPADE [42]. A very recent work, InfiniCity [34], follows a similar pipeline for 3D city generation. However, creating 3D cities presents greater complexity compared to 3D natural scenes. Buildings, as objects with the same semantic label, exhibit a wide range of appearances, unlike the relatively consistent appearance of objects like trees in natural scenes. This fact may decrease the quality of generated buildings when all

buildings in a city are given the same semantic label.

To handle the diversity of buildings in urban environments, we propose CityDreamer, a compositional generative model designed for unbounded 3D cities. As shown in Figure 2, CityDreamer differs from existing methods in that it splits the generation of building instances and background stuff like roads, green lands, and water areas into two separate modules: the building instance generator and the city background generator. Both generators adopt the bird’s eye view (BEV) scene representation and employ a volumetric renderer to generate photorealistic images via adversarial training. Notably, the scene parameterization is meticulously tailored to suit the distinct characteristics of background stuff and buildings. Background stuff in each category typically has similar appearances while exhibiting irregular textures. Hence, we introduce the generative hash grid to preserve naturalness while upholding 3D consistency. In contrast, building instances exhibit a wide range of appearances, but the texture of their façades often displays regular periodic patterns. Therefore, we design periodic positional encoding, which is simple yet effective for handling the diversity building façades. The compositor finally combines the rendered background stuff and building instances to generate a cohesive image.

To enhance the realism of our generated 3D cities, we construct a suite of CityGen Datasets, including OSM and GoogleEarth. The OSM dataset, sourced from OpenStreetMap [1], contains semantic maps and height fields of 80 cities, covering over 6,000 km². These maps show the locations of roads, buildings, green lands, and water areas, while the height fields primarily indicate building heights. The GoogleEarth dataset, gathered using Google Earth Studio [2], features 400 orbit trajectories in New York City. It includes 24,000 real-world city images, along with semantic and building instance segmentation. These annotations are automatically generated by projecting the 3D city layout, based on the OSM dataset, onto the images. The Google Earth dataset provides a wider variety of realistic urban images from different perspectives. Additionally, it can be easily expanded to include cities worldwide.

The contributions are summarized as follows:

- We propose CityDreamer, a compositional generative model designed specifically for unbounded 3D cities, which disentangles instance-oriented and stuff-oriented neural fields for buildings and backgrounds.
- We construct the CityGen Datasets, including OSM and GoogleEarth, with realistic city layouts and appearances, respectively. GoogleEarth includes images with multi-view consistency and building instance segmentation.
- The proposed CityDreamer showcases its superior capability in generating large-scale and diverse 3D cities. Additionally, it enables localized editing within the generated cities.

2. Related Work

3D-Aware GANs. Generative adversarial networks (GANs) [20] have achieved remarkable success in 2D image generation [27, 28]. Efforts to extend GANs into 3D space have also emerged, with some works [17, 40, 55] intuitively adopting voxel-based representations by extending the CNN backbone used in 2D. However, the high computational and memory cost of voxel grids and 3D convolution poses challenges in modeling unbounded 3D scenes. Recent advancements in neural radiance field (NeRF) [39] have led to the incorporation of volume rendering as a key inductive bias to make GANs 3D-aware. This enables GANs to learn 3D representations from 2D images [8, 18, 21, 41, 59]. Nevertheless, most of these methods are trained on curated datasets for bounded scenes, such as human faces [27], human bodies [25], and objects [56].

Scene-Level Content Generation. Unlike impressive 2D generative models that mainly target single categories or common objects, generating scene-level content is a challenging task due to the high diversity of scenes. Semantic image synthesis, such as [15, 22, 37, 42], shows promising results in generating scene-level content in the wild by conditioning on pixel-wise dense correspondence, such as semantic segmentation maps or depth maps. Some approaches have even achieved 3D-aware scene synthesis [22, 31, 36, 37, 50], but they may lack full 3D consistency or support feed-forward generation for novel worlds. Recent works like [7, 11, 34] have achieved 3D consistent scenes at infinity scale through unbounded layout extrapolation. Another bunch of work [3, 14, 43, 53] focus on indoor scene synthesis using expensive 3D datasets [13, 51] or CAD retrieval [16].

3. Our Approach

As shown in Figure 2, CityDreamer follows a four-step process to generate an unbounded 3D city. Initially, the unbounded layout generator (Sec. 3.1) creates an arbitrary large city layout L . Subsequently, the city background generator (Sec. 3.2) produces the background image \hat{I}_C along with its corresponding mask M_C . Next, the building instances generator (Sec. 3.3) generates images for building instances $\{\hat{I}_B^i\}_{i=1}^n$ and their respective masks $\{M_B^i\}_{i=1}^n$, where n is the number of building instances. Lastly, the compositor (Sec. 3.4) merges the rendered background and building instances into a single cohesive image I_C .

3.1. Unbounded City Layout Generator

City Layout Representation. The city layout determines the 3D objects present in the city and their respective locations. The objects can be categorized into six classes: roads, buildings, green lands, construction sites, water areas, and others. Moreover, there is an additional null class used to

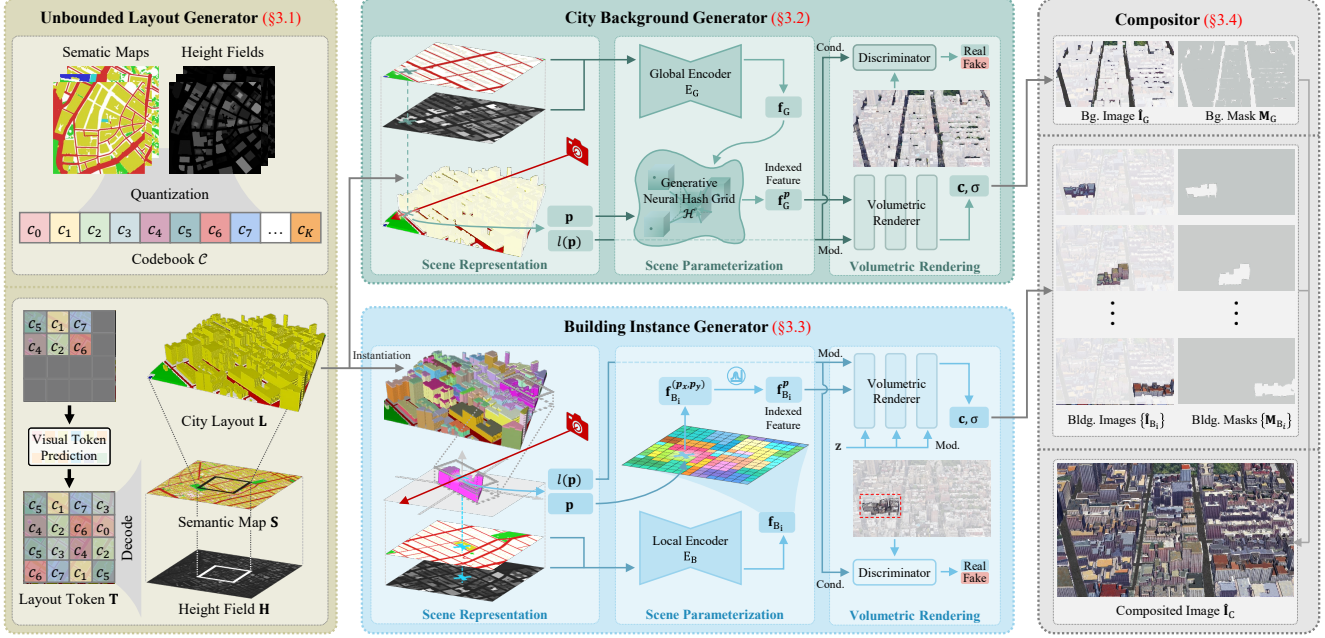


Figure 2. **Overview of CityDreamer.** The *unbounded layout generator* creates the city layout \mathbf{L} . Then, the *city background generator* performs ray-sampling to retrieve features from \mathbf{L} and generates the background image with a volumetric renderer focusing on background stuff like roads, green lands, and water areas. Similarly, the *building instance generator* renders the building instance image with another volumetric renderer. Finally, the *compositor* merges the rendered background and building images, producing a unified and coherent final image. Note that “Mod.,” “Cond.,” “Bg.,” and “Bldg.” denote “Modulation”, “Condition”, “Background”, and “Building”, respectively.

represent empty spaces in the 3D volumes. The city layout in CityDreamer, denoted as a 3D volume \mathbf{L} , is created by extruding the pixels in the semantic map \mathbf{S} based on the corresponding values in the height field \mathbf{H} . Specifically, the value of \mathbf{L} at (i, j, k) can be defined as

$$L_{(i,j,k)} = \begin{cases} S_{(i,j)} & \text{if } k \leq H_{(i,j)} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where 0 denotes empty spaces in the 3D volumes.

City Layout Generation. Obtaining unbounded city layouts is translated into generating extendable semantic maps and height fields. To this aim, we construct the unbounded layout generator based on MaskGIT [9], which inherently enables inpainting and extrapolation capabilities. Specifically, we employ VQVAE [47, 52] to tokenize the semantic map and height field patches, converting them into discrete latent space and creating a codebook $\mathcal{C} = \{c_k | c_k \in \mathbb{R}^D\}_{k=1}^K$. During inference, we generate the layout token \mathbf{T} in an autoregressive manner, and subsequently, we use the VQVAE’s decoder to generate a pair of semantic map \mathbf{S} and height field \mathbf{H} . Since VQVAE generates fixed-size semantic maps and height fields, we use image extrapolation to create arbitrary-sized ones. During this process, we adopt a sliding window to forecast a local layout token at every step, with a 25% overlap during the sliding.

Loss Functions. The VQVAE treats the generation of the height field and semantic map as two separate tasks, optimizing them using L1 Loss and Cross Entropy Loss \mathcal{E} , respectively. Additionally, to ensure sharpness in the height field around the edges of the buildings, we introduce an extra Smoothness Loss \mathcal{S} [38].

$$\ell_{\text{VQ}} = \lambda_{\text{R}} \|\hat{\mathbf{H}}_p - \mathbf{H}_p\| + \lambda_{\text{S}} \mathcal{S}(\hat{\mathbf{H}}_p, \mathbf{H}_p) + \lambda_{\text{E}} \mathcal{E}(\hat{\mathbf{S}}_p, \mathbf{S}_p) \quad (2)$$

where $\hat{\mathbf{H}}_p$ and $\hat{\mathbf{S}}_p$ denote the generated height field and semantic map patches, respectively. \mathbf{H}_p and \mathbf{S}_p are the corresponding ground truth. The autoregressive transformer in MaskGIT is trained using a reweighted ELBO loss [5].

3.2. City Background Generator

Scene Representation. Similar to SceneDreamer [11], we use the bird’s-eye-view (BEV) scene representation for its efficiency and expressive capabilities, making it easily applicable to unbounded scenes. Different from GAN-Craft [22] and InfiniCity [34], where features are parameterized to voxel corners, the BEV representation comprises a feature-free 3D volume generated from a height field and a semantic map, following Equation 1. Specifically, we initiate the process by selecting a local window with a resolution of $N_G^H \times N_G^W \times N_G^D$ from the city layout \mathbf{L} . This local window is denoted as $\mathbf{L}_G^{\text{Local}}$, which is generated from the corresponding height field $\mathbf{H}_G^{\text{Local}}$ and semantic map $\mathbf{S}_G^{\text{Local}}$.

Scene Parameterization. To achieve generalizable 3D representation learning across various scenes and align content with 3D semantics, it is necessary to parameterize the scene representation into a latent space, making adversarial learning easier. For background stuff, we adopt the generative neural hash grid [11] to learn generalizable features across scenes by modeling the hyperspace beyond 3D space. Specifically, we first encode the local scene $(\mathbf{H}_G^{\text{Local}}, \mathbf{S}_G^{\text{Local}})$ using the global encoder E_G to produce the compact scene-level feature $\mathbf{f}_G \in \mathbb{R}^{d_G}$.

$$\mathbf{f}_G = E_G(\mathbf{H}_G^{\text{Local}}, \mathbf{S}_G^{\text{Local}}) \quad (3)$$

By leveraging a learnable neural hash function \mathcal{H} , the indexed feature $\mathbf{f}_G^{\mathbf{p}}$ at 3D position $\mathbf{p} \in \mathbb{R}^3$ can be obtained by mapping \mathbf{p} and \mathbf{f}_G into a hyperspace, *i.e.*, $\mathbb{R}^{3+d_G} \rightarrow \mathbb{R}^{N_G^C}$.

$$\mathbf{f}_G^{\mathbf{p}} = \mathcal{H}(\mathbf{p}, \mathbf{f}_G) = \left(\bigoplus_{i=1}^{d_G} f_G^i \pi^i \bigoplus_{j=1}^3 p^j \pi^j \right) \bmod T \quad (4)$$

where \oplus denotes the bit-wise XOR operation. π^i and π^j represent large and unique prime numbers. We construct N_H^L levels of multi-resolution hash grids to represent multi-scale features, T is the maximum number of entries per level, and N_G^C denotes the number of channels in each unique feature vector.

Volumetric Rendering. In a perspective camera model, each pixel in the image corresponds to a camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{v}$, where the ray originates from the center of projection \mathbf{o} and extends in the direction \mathbf{v} . Thus, the corresponding pixel value $C(\mathbf{r})$ is derived from an integral.

$$C(\mathbf{r}) = \int_0^\infty T(t) \mathbf{c}(\mathbf{f}_G^{\mathbf{r}(t)}, l(\mathbf{r}(t))) \boldsymbol{\sigma}(\mathbf{f}_G^{\mathbf{r}(t)}) dt \quad (5)$$

where $T(t) = \exp(-\int_0^t \sigma(\mathbf{f}_G^{\mathbf{r}(s)}) ds)$. $l(\mathbf{p})$ represent the semantic label at the 3D position \mathbf{p} . \mathbf{c} and $\boldsymbol{\sigma}$ denote the color and volume density, respectively.

Loss Function. The city background generator is trained using a hybrid objective, which includes a combination of a reconstruction loss and an adversarial learning loss. Specifically, we leverage the L1 loss, perceptual loss \mathcal{P} [26], and GAN loss \mathcal{G} [32] in this combination.

$$\ell_G = \lambda_{L1} \|\hat{\mathbf{I}}_G - \mathbf{I}_G\| + \lambda_P \mathcal{P}(\hat{\mathbf{I}}_G, \mathbf{I}_G) + \lambda_G \mathcal{G}(\hat{\mathbf{I}}_G, \mathbf{S}_G) \quad (6)$$

where \mathbf{I}_G denotes the ground truth background image. \mathbf{S}_G is the semantic map in perspective view generated by accumulating semantic labels sampled from the $\mathbf{L}_G^{\text{Local}}$ along each ray. The weights of the three losses are denoted as λ_{L1} , λ_P , and λ_G . Note that ℓ_G is solely applied to pixels with semantic labels belonging to background stuff.

3.3. Building Instance Generator

Scene Representation. Just like the city background generator, the building instance generator also uses the BEV scene representation. In the building instance generator, we extract a local window denoted as $\mathbf{L}_{B_i}^{\text{Local}}$ from the city layout \mathbf{L} , with a resolution of $N_B^H \times N_B^W \times N_B^D$, centered around the 2D center $(c_{B_i}^x, c_{B_i}^y)$ of the building instance B_i . The height field and semantic map used to generate $\mathbf{L}_{B_i}^{\text{Local}}$ can be denoted as $\mathbf{H}_{B_i}^{\text{Local}}$ and $\mathbf{S}_{B_i}^{\text{Local}}$, respectively. As all buildings have the same semantic label in \mathbf{S} , we perform building instantiation by detecting connected components. We observe that the façades and roofs of buildings in real-world scenes exhibit distinct distributions. Consequently, we assign different semantic labels to the façade and roof of the building instance B_i in $\mathbf{L}_{B_i}^{\text{Local}}$, with the top-most voxel layer being assigned the roof label. The rest building instances are omitted in $\mathbf{L}_{B_i}^{\text{Local}}$ by assigned with the null class.

Scene Parameterization. In contrast to the city background generator, the building instance generator employs a novel scene parameterization that relies on pixel-level features generated by a local encoder E_B . Specifically, we start by encoding the local scene $(\mathbf{H}_{B_i}^{\text{Local}}, \mathbf{S}_{B_i}^{\text{Local}})$ using E_B , resulting in the pixel-level feature \mathbf{f}_{B_i} , which has a resolution of $N_B^H \times N_B^W \times N_B^C$.

$$\mathbf{f}_{B_i} = E_B(\mathbf{H}_{B_i}^{\text{Local}}, \mathbf{S}_{B_i}^{\text{Local}}) \quad (7)$$

Given a 3D position $\mathbf{p} = (p_x, p_y, p_z)$, the corresponding indexed feature $\mathbf{f}_{B_i}^{\mathbf{p}}$ can be computed as

$$\mathbf{f}_{B_i}^{\mathbf{p}} = \mathcal{O}(\text{Concat}(\mathbf{f}_{B_i}^{(p_x, p_y)}, p_z)) \quad (8)$$

where $\text{Concat}(\cdot)$ is the concatenation operation. $\mathbf{f}_{B_i}^{(p_x, p_y)} \in \mathbb{R}^{N_B^C}$ denotes the feature vector at (p_x, p_y) . $\mathcal{O}(\cdot)$ is the positional encoding function used in the vanilla NeRF [39].

$$\mathcal{O}(x) = \{\sin(2^i \pi x), \cos(2^i \pi x)\}_{i=0}^{N_P^L-1} \quad (9)$$

Note that $\mathcal{O}(\cdot)$ is applied individually to each value in the given feature x , which are normalized to lie within the range of $[-1, 1]$.

Volumetric Rendering. Different from the volumetric rendering used in the city background generator, we incorporate a style code \mathbf{z} in the building instance generator to capture the diversity of buildings. The corresponding pixel value $C(\mathbf{r})$ is obtained through an integral process.

$$C(\mathbf{r}) = \int_0^\infty T(t) \mathbf{c}(\mathbf{f}_{B_i}^{\mathbf{r}(t)}, \mathbf{z}, l(\mathbf{r}(t))) \boldsymbol{\sigma}(\mathbf{f}_{B_i}^{\mathbf{r}(t)}) dt \quad (10)$$

Note that the camera ray $\mathbf{r}(t)$ is normalized with respect to $(c_{B_i}^x, c_{B_i}^y, 0)$ as the origin.

Loss Function. For training the building instance generator, we exclusively use the GAN Loss. Mathematically, it

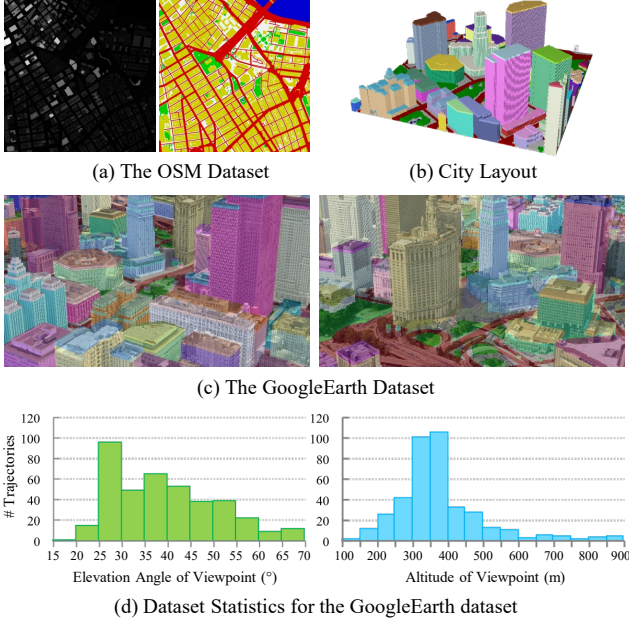


Figure 3. **Overview of CityGen Datasets.** (a) The OSM dataset comprising paired height fields and semantic maps provides real-world city layouts. (b) The city layout, generated from the height field and semantic map, facilitates automatic annotation generation. (c) The GoogleEarth dataset includes real-world city appearances alongside semantic segmentation and building instance segmentation. (d) The dataset statistics demonstrate the variety of perspectives available in the GoogleEarth dataset.

can be represented as

$$\ell_B = \mathcal{G}(\hat{\mathbf{I}}_{B_i}, \mathbf{S}_{B_i}) \quad (11)$$

where \mathbf{S}_{B_i} denotes the semantic map of building instance B_i in perspective view, which is generated in a similar manner to \mathbf{S}_G . Note that ℓ_B is exclusively applied to pixels with semantic labels belonging to the building instance.

3.4. Composer

Since there are no corresponding ground truth images for the images generated by the City Background Generator and Building Instance Generator, it is not possible to train neural networks to merge these images. Therefore, the compositor uses the generated images $\hat{\mathbf{I}}_G$ and $\{\hat{\mathbf{I}}_{B_i}\}_{i=1}^n$, along with their corresponding binary masks \mathbf{M}_G and $\{\mathbf{M}_{B_i}\}_{i=1}^n$, the compositor combines them into a unified image \mathbf{I}_C , which can be represented as

$$\mathbf{I}_C = \hat{\mathbf{I}}_G \mathbf{M}_G + \sum_{i=1}^n \hat{\mathbf{I}}_{B_i} \mathbf{M}_{B_i} \quad (12)$$

where n is the number of building instances.

Table 1. **A Comparison of GoogleEarth with representative city-related datasets.** Note that the number of images and area are counted based on real-world images. “sate.” represents satellite. “inst.,” “sem.,” and “plane” denote “instance segmentation,” “semantic segmentation,” and “plane segmentation” respectively.

Dataset	# Images	Area	View	Annotation	3D
KITTI [19]	15 k	-	street	sem.	✗
Cityscapes [12]	25 k	-	street	sem.	✗
SpaceNet MOVI [54]	6.0 k	-	sate.	inst.	✗
OmniCity [30]	108 k	-	street/sate.	inst./plane	✗
HoliCity [61]	6.3 k	20 km ²	street	sem./plane	✓
UrbanScene3D [35]	6.1 k	3 km ²	drone	inst.	✓
GoogleEarth	24 k	25 km ²	drone	inst./sem.	✓

4. CityGen Datasets

The OSM Dataset. The OSM dataset, sourced from OpenStreetMap [1], is composed of the rasterized semantic maps and height fields of 80 cities worldwide, spanning an area of more than 6,000 km². During the rasterization process, vectorized geometry information is converted into images by translating longitude and latitude into the EPSG:3857 coordinate system at zoom level 18, approximately 0.597 meters per pixel. As shown in Figure 3(a), the segmentation maps use red, yellow, green, cyan, and blue colors to denote the positions of roads, buildings, green lands, construction sites, and water areas, respectively. The height fields primarily represent the height of buildings, with their values derived from OpenStreetMap. For roads, the height values are set to 4, while for water areas, they are set to 0. Additionally, the height values for trees are sampled from perlin noise [44], ranging from 8 to 16.

The GoogleEarth Dataset. The GoogleEarth dataset is collected from Google Earth Studio [2], including 400 orbit trajectories in Manhattan and Brooklyn. Each trajectory consists of 60 images, with orbit radiuses ranging from 125 to 813 meters and altitudes varying from 112 to 884 meters. In addition to the images, Google Earth Studio provides camera intrinsic and extrinsic parameters, making it possible to create automated annotations for semantic and building instance segmentation. Specifically, for building instance segmentation, we initially perform connected components detection on the semantic maps to identify individual building instances. Then, the city layout is created following Equation 1, as demonstrated in Figure 3(b). Finally, the annotations are generated by projecting the city layout onto the images, using the camera parameters, as shown in Figure 3(c). Table 1 presents a comparative overview between GoogleEarth and other datasets related to urban environments. Among datasets that offer 3D models, GoogleEarth stands out for its extensive coverage of real-world images, encompassing the largest area, and providing annotations for both semantic and instance segmentation. Figure 3(d) offers an analysis of viewpoint altitudes

and elevations in the GoogleEarth dataset, highlighting its diverse camera viewpoints. This diversity enhances neural networks’ ability to generate cities from a broader range of perspectives. Additionally, leveraging Google Earth and OpenStreetMap data allows us to effortlessly expand our dataset to encompass more cities worldwide.

5. Experiments

5.1. Evaluation Protocols

During evaluation, we use the Unbounded Layout Generator to generate 1024 distinct city layouts. For each scene, we sample 20 different styles by randomizing the style code \mathbf{z} . Each sample is transformed into a fly-through video consisting of 40 frames, each with a resolution of 960×540 pixels and any possible camera trajectory. Subsequently, we randomly select frames from these video sequences for evaluation. The evaluation metrics are as follows:

FID and KID. Fréchet Inception Distance (FID) [23] and Kernel Inception Distance (KID) [4] are metrics for the quality of generated images. We compute FID and KID between a set of 15,000 generated frames and an evaluation set comprising 15,000 images randomly sampled from the GoogleEarth dataset.

Depth Error. We employ depth error (DE) to assess the 3D geometry, following a similar approach to EG3D [8]. Using a pre-trained model [46], we generate pseudo ground truth depth maps for generated frames by accumulating density σ . Both the “ground truth” depth and the predicted depth are normalized to zero mean and unit variance to eliminate scale ambiguity. DE is computed as the L2 distance between the two normalized depth maps. We assess this depth error on 100 frames for each evaluated method.

Camera Error. Following SceneDreamer [11], we introduce camera error (CE) to assess multi-view consistency. CE quantifies the difference between the inference camera trajectory and the estimated camera trajectory from COLMAP [48]. It is calculated as the scale-invariant normalized L2 distance between the reconstructed and generated camera poses.

5.2. Implementation Details

Hyperparameters:

Unbounded Layout Generator. The codebook size K is set to 512, and each code’s dimension D is set to 512. The height field and semantic map patches are cropped to a size of 512×512 , and compressed by a factor of 16. The loss weights, λ_R , λ_S , and λ_E , are 10, 10, 1, respectively.

City Background Generator. The local window resolution N_G^H , N_G^W , and N_G^D are set to 1536, 1536, and 640, respectively. The dimension of the scene-level features d_G is 2. For the generative hash grid, we use $N_H^L = 16$, $T = 2^{19}$, and $N_G^C = 8$. The unique prime numbers in Equation 4

Table 2. **Quantitative comparison.** The best values are highlighted in bold. Note that the results of InfiniCity are not included in this comparison as it is not open-sourced.

Methods	FID ↓	KID ↓	DE ↓	CE ↓
SGAM [49]	277.64	0.358	0.575	239.291
PersistentNature [7]	123.83	0.109	0.326	86.371
SceneDreamer [11]	213.56	0.216	0.152	0.186
CityDreamer	97.38	0.096	0.147	0.060

are set to $\pi^1 = 1$, $\pi^2 = 2654435761$, $\pi^3 = 805459861$, $\pi^4 = 3674653429$, and $\pi^5 = 2097192037$. The loss function weights, λ_{L1} , λ_P , and λ_G , are 10, 10, 0.5, respectively. *Building Instance Generator.* The local window resolution N_B^H , N_B^W , and N_B^D are set to 672, 672, and 640, respectively. The number of channels N_B^C of the pixel-level features is 63. The dimension N_P^L is set to 10.

Training Details:

Unbounded Layout Generator. The VQVAE is trained with a batch size of 16 using an Adam optimizer with $\beta = (0.5, 0.9)$ and a learning rate of 7.2×10^{-5} for 1,250,000 iterations. The autoregressive transformer is trained with a batch size of 80 using an Adam optimizer with $\beta = (0.9, 0.999)$ and a learning rate of 2×10^{-4} for 250,000 iterations.

City Background and Building Instance Generators. Both generators are trained using an Adam optimizer with $\beta = (0, 0.999)$ and a learning rate of 10^{-4} . The discriminators are optimized using an Adam optimizer with $\beta = (0, 0.999)$ and a learning rate of 10^{-5} . The training lasts for 298,500 iterations with a batch size of 8. The images are randomly cropped to a size of 192×192 .

5.3. Main Results

Comparison Methods. We compare CityDreamer against four state-of-the-art methods: SGAM [49], PersistentNature [7], SceneDreamer [11], and InfiniCity [34]. With the exception of InfiniCity, whose code is not available, the remaining methods are retrained using the released code on the GoogleEarth dataset to ensure a fair comparison. SceneDreamer initially uses simplex noise for layout generation, which is not ideal for cities, so it is replaced with the unbounded layout generator from CityDreamer.

Qualitative Comparison. Figure 4 provides qualitative comparisons against baselines. SGAM struggles to produce realistic results and maintain good 3D consistency because extrapolating views for complex 3D cities can be extremely challenging. PersistentNature employs tri-plane representation, but it encounters challenges in generating realistic renderings. SceneDreamer and InfiniCity both utilize voxel grids as their representation, but they still suffer from severe structural distortions in buildings because all buildings are given the same semantic label. In comparison, the proposed CityDreamer generates more realistic and diverse re-

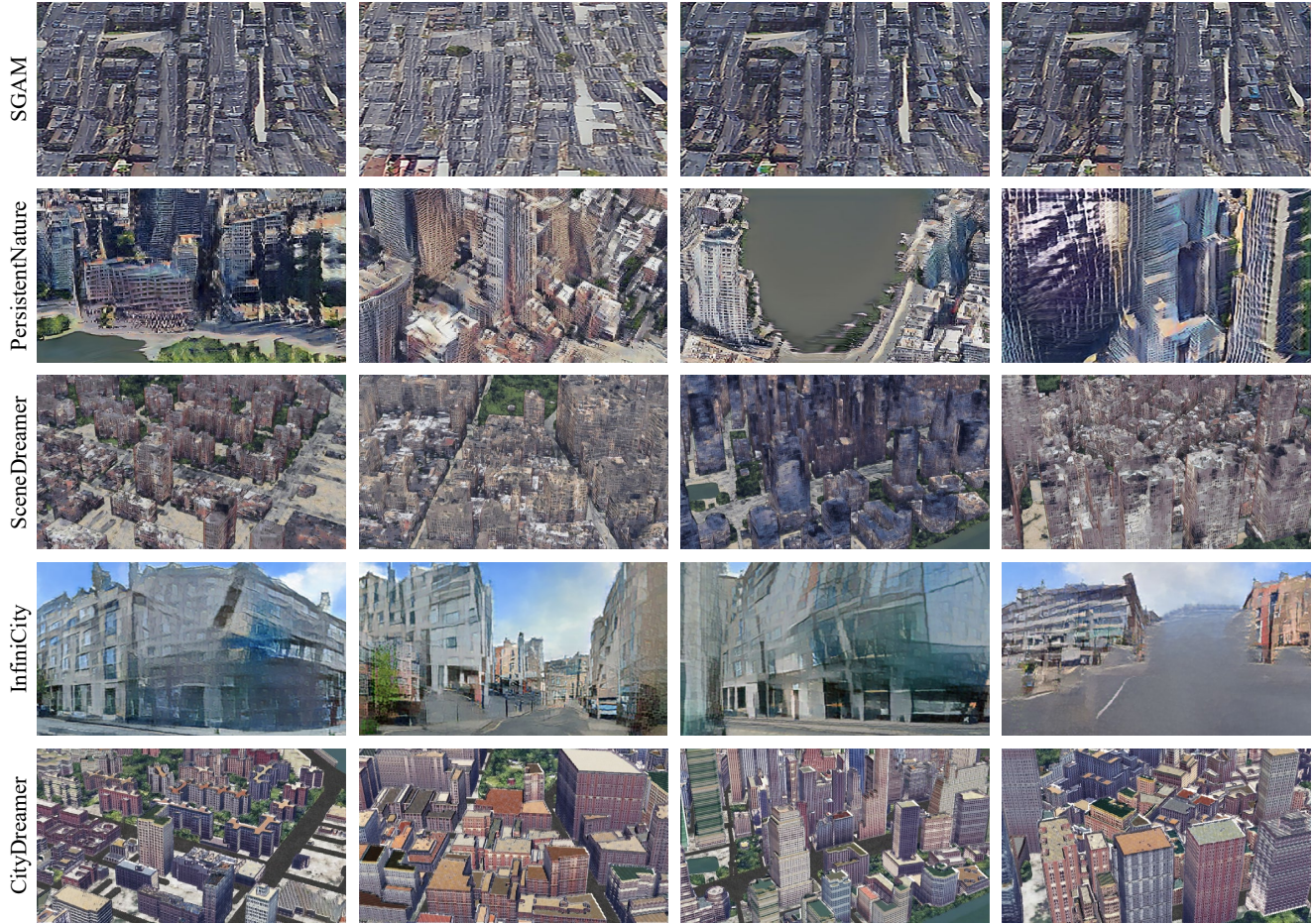


Figure 4. **Qualitative comparison.** The proposed CityDreamer produces more realistic and diverse results compared to all baselines. Note that the visual results of InfiniCity [34] are provided by the authors and zoomed for optimal viewing.

sults compared to all the baselines.

Quantitative Comparison. Table 2 presents the quantitative metrics of the proposed approach compared to the baselines. CityDreamer exhibits significant improvements in FID and KID, which is consistent with the visual comparisons. Moreover, CityDreamer demonstrates the capability to maintain accurate 3D geometry and view consistency while generating photorealistic images, as evident by the lowest DE and CE errors compared to the baselines.

User Study. To better assess the 3D consistency and quality of the unbounded 3D city generation, we conduct an output evaluation [6] as the user study. In this survey, we ask 20 volunteers to rate each generated camera trajectory based on three aspects: 1) the perceptual quality of the imagery, 2) the level of 3D realism, and 3) the 3D view consistency. The scores are on a scale of 1 to 5, with 5 representing the best rating. The results are presented in Figure 5, showing that the proposed method significantly outperforms the baselines by a large margin.

5.4. Ablation Study

Effectiveness of Unbounded Layout Generator. The Unbounded Layout Generator plays a critical role in generating “unbounded” city layouts. We compare it with InfinityGAN [33] used in InfiniCity and a rule-based city layout generation method, IPSM [10], as shown in Table 4. Following InfiniCity [34], we use FID and KID to evaluate the quality of the generated layouts. Compared to IPSM and InfinityGAN, Unbounded Layout Generator achieves better results in terms of all metrics. The qualitative results shown in Figure I in the supplementary material also demonstrate the effectiveness of the proposed method.

Effectiveness of Building Instance Generator. We emphasize the crucial role of the building instance generator in the success of unbounded 3D city generation. To demonstrate its effectiveness, we conducted an ablation study on the building instance generator. We compared two optional designs: (1) Removing the building instance generator from CityDreamer, *i.e.*, the model falling back to SceneDreamer.

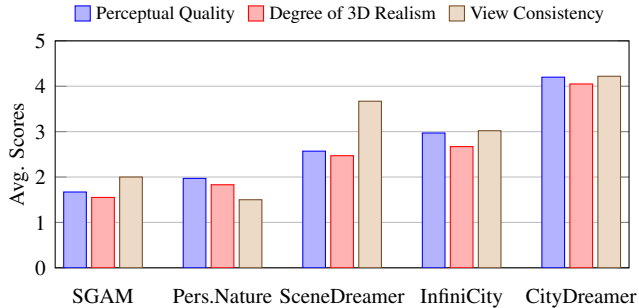


Figure 5. **User study on unbounded 3D city generation.** All scores are in the range of 5, with 5 indicating the best. Note that “Pers.Nature” denotes “PersistentNature” [7].

(2) All buildings are generated at once by the building instance generator, without providing any instance labels. The quantitative results presented in Table 4 demonstrate the effectiveness of both the instance labels and the building instance generator. Refer to Figure II in the supplementary material for more qualitative comparisons.

Effectiveness of Scene Parameterization. Scene parameterization directly impacts the quality of 3D city generation. The city background generator utilizes HashGrid with patch-wise features from the global encoder, while the building instance generator uses vanilla SinCos positional encoding with pixel-wise features from the local encoder. We compare different scene parameterizations in both the city background generator and the building instance generator. Table 5 shows that using local encoders in background generation or using global encoders in building generation leads to considerable degradation in image quality, indicated by poor metrics. According to Equation 4, the output of HashGrid is determined by the scene-level features and 3D position. While HashGrid enhances the multi-view consistency of the generated background, it also introduces challenges in building generation, leading to less structurally reasonable buildings. In contrast, the inherent periodicity of SinCos makes it easier for the network to learn the periodicity of building façades, leading to improved results in building generation. Refer to Sec. A.2 in the supplementary material for a detailed discussion.

5.5. Further Discussions

Applications. This research primarily benefits applications that require efficient content creation, with notable examples being the entertainment industry. There is a strong demand to generate content for computer games and movies within this field.

Limitations. 1) The generation of the city layout involves raising voxels to a specific height, which means that concave geometries like caves and tunnels cannot be modeled and generated. 2) During the inference process, the buildings are generated individually, resulting in a slightly higher

Table 3. **Effectiveness of Unbounded Layout Generator.** The best values are highlighted in bold. The images are centrally cropped to a size of 4096×4096.

Methods	FID ↓	KID ↓
IPSM [10]	321.47	0.502
InfinityGAN [33]	183.14	0.288
Ours	124.45	0.123

Table 4. **Effectiveness of Building Instance Generator.** The best values are highlighted in bold. Note that “w/o BIG.” indicates the removal of Building Instance Generator from CityDreamer. “w/o Ins.” denotes the absence of building instance labels in the building instance generator.

Methods	FID ↓	KID ↓	DE ↓	CE ↓
w/o BIG.	213.56	0.216	0.152	0.186
w/o Ins.	117.75	0.124	0.148	0.098
Ours	97.38	0.096	0.147	0.060

Table 5. **Effectiveness of different generative scene parameterization.** The best values are highlighted in bold. Note that “CBG.” and “BIG.” denote City Background Generator and Building Instance Generator, respectively. “Enc.” and “P.E.” represent “Encoder” and “Positional Encoding”, respectively.

CBG.		BIG.		FID ↓	KID ↓	DE ↓	CE ↓
Enc.	P.E.	Enc.	P.E.				
Local	SinCos	Global	Hash	219.30	0.233	0.154	0.452
Local	SinCos	Local	SinCos	107.63	0.125	0.149	0.078
Global	Hash	Global	Hash	213.56	0.216	0.153	0.186
Global	Hash	Local	SinCos	97.38	0.096	0.147	0.060

computation cost. Exploring ways to reduce the inference cost would be beneficial for future work.

6. Conclusion

In this paper, we propose CityDreamer, a compositional generative model designed specifically for unbounded 3D cities. Compared to existing methods that treat buildings as a single class of objects, CityDreamer separates the generation of building instances from background stuff, allowing for better handling of the diverse appearances of buildings. Additionally, we create a suite of CityGen Datasets, including OSM and GoogleEarth, providing more realistic city layouts and appearances, and easily scalable to include other cities worldwide. CityDreamer achieves state-of-the-art performance not only in generating realistic 3D cities but also in localized editing within the generated cities.

Acknowledgments This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20221-0012), NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

- [1] <https://openstreetmap.org>. 2, 5
- [2] <https://earth.google.com/studio>. 2, 5
- [3] Miguel Ángel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, Afshin Dehghan, and Joshua M. Susskind. GAUDI: A neural architect for immersive 3d scene generation. In *NeurIPS*, 2022. 2
- [4] Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018. 6
- [5] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P. Breckon, and Chris G. Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *ECCV*, 2022. 3
- [6] Zoya Bylinskii, Laura Mariah Herman, Aaron Hertzmann, Stefanie Hutka, and Yile Zhang. Towards better user studies in computer graphics and vision. *Foundations and Trends in Computer Graphics and Vision*, 15(3):201–252, 2023. 7
- [7] Lucy Chai, Richard Tucker, Zhengqi Li, Phillip Isola, and Noah Snively. Persistent Nature: A generative model of unbounded 3D worlds. In *CVPR*, 2023. 1, 2, 6, 8
- [8] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 2, 6
- [9] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *CVPR*, 2022. 3
- [10] Guoning Chen, Gregory Esch, Peter Wonka, Pascal Müller, and Eugene Zhang. Interactive procedural street modeling. *ACM TOG*, 27(3):103, 2008. 7, 8
- [11] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. SceneDreamer: Unbounded 3D scene generation from 2D image collections. *TPAMI*, 45(12):15562–15576, 2023. 1, 2, 3, 4, 6
- [12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 5
- [13] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017. 2
- [14] Terrance DeVries, Miguel Ángel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *ICCV*, 2021. 2
- [15] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2
- [16] Huan Fu, Bowen Cai, Lin Gao, Lingxiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3D furnished rooms with layouts and semantics. In *ICCV*, 2021. 2
- [17] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2D views of multiple objects. In *3DV*, 2017. 2
- [18] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. GET3D: A generative model of high quality 3D textured shapes learned from images. In *NeurIPS*, 2022. 2
- [19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 5
- [20] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2
- [21] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. StyleNeRF: A style-based 3D aware generator for high-resolution image synthesis. In *ICLR*, 2022. 2
- [22] Zekun Hao, Arun Mallya, Serge J. Belongie, and Ming-Yu Liu. GANCraft: Unsupervised 3D neural rendering of minecraft worlds. In *ICCV*, 2021. 1, 2, 3
- [23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 6
- [24] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. EVA3D: compositional 3D human generation from 2D image collections. In *ICLR*, 2023. 1
- [25] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE TPAMI*, 36(7):1325–1339, 2014. 2
- [26] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 4
- [27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE TPAMI*, 43(12):4217–4228, 2021. 2
- [28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2
- [29] Nikos Kolotouros, Thiemo Ailidieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. DreamHuman: Animatable 3D avatars from text. *arXiv*, 2306.09329, 2023. 1
- [30] Weijia Li, Yawen Lai, Linning Xu, Yuanbo Xiangli, Jinhua Yu, Conghui He, Gui-Song Xia, and Dahua Lin. OmniCity: Omnipotent city understanding with multi-level and multi-view images. In *CVPR*, 2023. 5
- [31] Zhengqi Li, Qianqian Wang, Noah Snively, and Angjoo Kanazawa. InfiniteNature-Zero: Learning perpetual view generation of natural scenes from single images. In *ECCV*, 2022. 2
- [32] Jae Hyun Lim and Jong Chul Ye. Geometric GAN. *arXiv*, 1705.02894, 2017. 4
- [33] Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, and Ming-Hsuan Yang. InfinityGan: Towards infinite-pixel image synthesis. In *ICLR*, 2022. 7, 8
- [34] Chieh Hubert Lin, Hsin-Ying Lee, Willi Menapace, Menglei Chai, Aliaksandr Siarohin, Ming-Hsuan Yang, and Sergey

- Tulyakov. InfiniCity: Infinite-scale city synthesis. In *ICCV*, 2023. 1, 2, 3, 6, 7
- [35] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: The urbanscene3d dataset. In *ECCV*, 2022. 5
- [36] Andrew Liu, Ameesh Makadia, Richard Tucker, Noah Snavely, Varun Jampani, and Angjoo Kanazawa. Infinite Nature: Perpetual view generation of natural scenes from a single image. In *ICCV*, 2021. 2
- [37] Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-consistent video-to-video synthesis. In *ECCV*, 2020. 2
- [38] Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Un-supervised learning of optical flow with a bidirectional census loss. In *AAAI*, 2018. 3
- [39] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 4
- [40] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3D representations from natural images. In *CVPR*, 2019. 2
- [41] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-resolution 3D-consistent image and geometry generation. In *CVPR*, 2022. 2
- [42] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 1, 2
- [43] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. ATISS: autoregressive transformers for indoor scene synthesis. In Marc’Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *NeurIPS*, 2021. 2
- [44] Ken Perlin. An image synthesizer. In *SIGGRAPH*, 1985. 5
- [45] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan T. Barron, Yuanzhen Li, and Varun Jampani. DreamBooth3D: Subject-driven text-to-3D generation. *arXiv*, 2303.13508, 2023. 1
- [46] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3):1623–1637, 2022. 6
- [47] Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *NeurIPS*, 2019. 3
- [48] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 6
- [49] Yuan Shen, Wei-Chiu Ma, and Shenlong Wang. SGAM: building a virtual 3D world through simultaneous generation and mapping. In *NeurIPS*, 2022. 6
- [50] Zifan Shi, Yujun Shen, Jiapeng Zhu, Dit-Yan Yeung, and Qifeng Chen. 3D-aware indoor scene synthesis with depth priors. In *ECCV*, 2022. 2
- [51] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Yuheng Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard A. Newcombe. The Replica Dataset: A digital replica of indoor spaces. *arXiv*, 1906.05797, 2019. 2
- [52] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NIPS*, 2017. 3
- [53] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. Sceneformer: Indoor scene generation with transformers. In *3DV*, 2021. 2
- [54] Nicholas Weir, David Lindenbaum, Alexei Bastidas, Adam Van Etten, Varun Kumar Vijay, Sean McPherson, Jacob Shermeyer, and Hanlin Tang. Spacenet MVOI: A multi-view overhead imagery dataset. In *ICCV*, 2019. 5
- [55] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *NIPS*, 2016. 2
- [56] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. OmniObject3D: Large-vocabulary 3D object dataset for realistic perception, reconstruction and generation. In *CVPR*, 2023. 2
- [57] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2Vox: Context-aware 3D reconstruction from single and multi-view images. In *ICCV*, 2019. 1
- [58] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2Vox++: Multi-scale context-aware 3D object reconstruction from single and multiple images. *IJCV*, 128(12):2919–2935, 2020. 1
- [59] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. GIRAFFE HD: A high-resolution 3D-aware generative model. In *CVPR*, 2022. 2
- [60] Chi Zhang, Yiwen Chen, Yijun Fu, Zhenglin Zhou, Gang Yu, Billz Wang, Bin Fu, Tao Chen, Guosheng Lin, and Chunhua Shen. StyleAvatar3D: Leveraging image-text diffusion models for high-fidelity 3D avatar generation. *arXiv*, 2305.19012, 2023. 1
- [61] Yichao Zhou, Jingwei Huang, Xili Dai, Linjie Luo, Zhili Chen, and Yi Ma. HoliCity: A city-scale data platform for learning holistic 3D structures. *arXiv*, 2008.03286, 2020. 5