UrbanGen: Urban Generation with Compositional and Controllable Neural Fields

Yuanbo Yang, Yujun Shen, Yue Wang, Andreas Geiger, Yiyi Liao ✉

Abstract—Despite the rapid progress in generative radiance fields, most existing methods focus on object-centric applications and are not able to generate complex urban scenes. In this paper, we propose UrbanGen, a solution for the challenging task of generating urban radiance fields with photorealistic rendering, accurate geometry, high controllability, and diverse city styles. Our key idea is to leverage a coarse 3D panoptic prior, represented by a semantic voxel grid for stuff and bounding boxes for countable objects, to condition a compositional generative radiance field. This panoptic prior simplifies the task of learning complex urban geometry, enables disentanglement of stuff and objects, and provides versatile control over both. Moreover, by combining semantic and geometry losses with adversarial training, our method faithfully adheres to the input conditions, allowing for joint rendering of semantic and depth maps alongside RGB images. In addition, we collect a unified dataset with images and their panoptic priors in the same format from 3 diverse real-world datasets: KITTI-360, nuScenes, and Waymo, and train a city style-aware model on this data. Our systematic study shows that UrbanGen outperforms state-of-the-art generative radiance field baselines in terms of image fidelity and geometry accuracy for urban scene generation. Furthermore, UrbenGen brings a new set of controllability features, including large camera movements, stuff editing, and city style control.

Index Terms—Urban Scenes, Generative Radiance Fields, 3D GANs, Neural Rendering

1 Introduction

This paper focuses on advancing generative radiance fields of urban scenes which will enable many important applications, e.g., serving as neural simulators for autonomous driving [12] or generating scenes for the gaming industry. In addition to rendering photorealistic images, an ideal generative urban radiance field should be capable of: i) learning accurate geometry, ii) providing high controllability, and iii) synthesizing different city styles. Although remarkable advances have been made in generative radiance fields using adversarial training [54], existing methods struggle to fulfill the above requirements. But what makes generative urban radiance fields so challenging?

Firstly, most existing 3D generative radiance field methods focus on "alignable" object-centric scenes like Carla Cars [17] and human faces [27]. This simplifies the task, as the generative model can learn the underlying alignment in canonical space, hence minimizing the residual geometry and appearance variations that need to be captured. Urban scenes, however, pose a great challenge as they lack a shared canonical geometry like object-centric scenarios such as human faces. Secondly, many properties of existing generative radiance field models, including semantics and appearance, remain entangled [60], which greatly limits the controllability of the generated content. To address these two challenges, GIRAFFE [43] and DiscoScene [67] propose to learn foreground cars in a canonical space from urban scene images using object layouts, learning plausible car geometry and enabling control over poses of cars. However, they generate the scene background as a simple 2D image, resulting in limited control over camera pose and semantic editing. Lastly, synthesizing different city styles requires training generative radiance fields on different urban datasets, which is non-trivial due to the diversity in scale of existing urban datasets. A dataset of small scale, e.g., nuScenes [8] with only 1,000 20-second video clips, may lead to unstable training and hence unsatisfying results.

In this work, we propose UrbanGen, a generative radiance field of urban scenes aiming to fulfill the requirements of photorealistic rendering, plausible geometry generation, high controllability, and diversity in city style. Our key idea is to leverage a 3D urban *panoptic prior*, formatted as a semantic voxel grid for *stuff* and bounding boxes for *countable objects*, to condition generative radiance fields for compositional and controllable generation. As we demonstrate in our experiments, the coarse geometry information captured by this prior eases the task of geometry generation through 2D supervision. Moreover, it seamlessly enables versatile control over the stuff and objects by editing the panoptic prior. We further collect a diverse set of panoptic priors from various urban datasets and utilize them for joint training. This allows for utilizing cross-dataset information and improves the performance compared to training solely on each individual dataset, thereby enabling synthesizing diverse city styles and even applying the style of one city to the layout of another.

Our model represents the scene as a compositional neural radiance field consisting of stuff, objects, and background. We inject our panoptic prior into the generative model in two ways: the semantic voxel grid is used to con-

- *Y. Yang, Y. Wang, and Y. Liao are affiliated with Zhejiang University, China.*
- *Y. Shen is affiliated with Ant Group, China.*
- *A. Geiger is affiliated with the Autonomous Vision Group (AVG) at the University of Tübingen and Tübingen AI Center, Germany.*
- ✉*Corresponding author: yiyi.liao@zju.edu.cn.*

- *Project Page: https://xdimlab.github.io/UrbanGen*

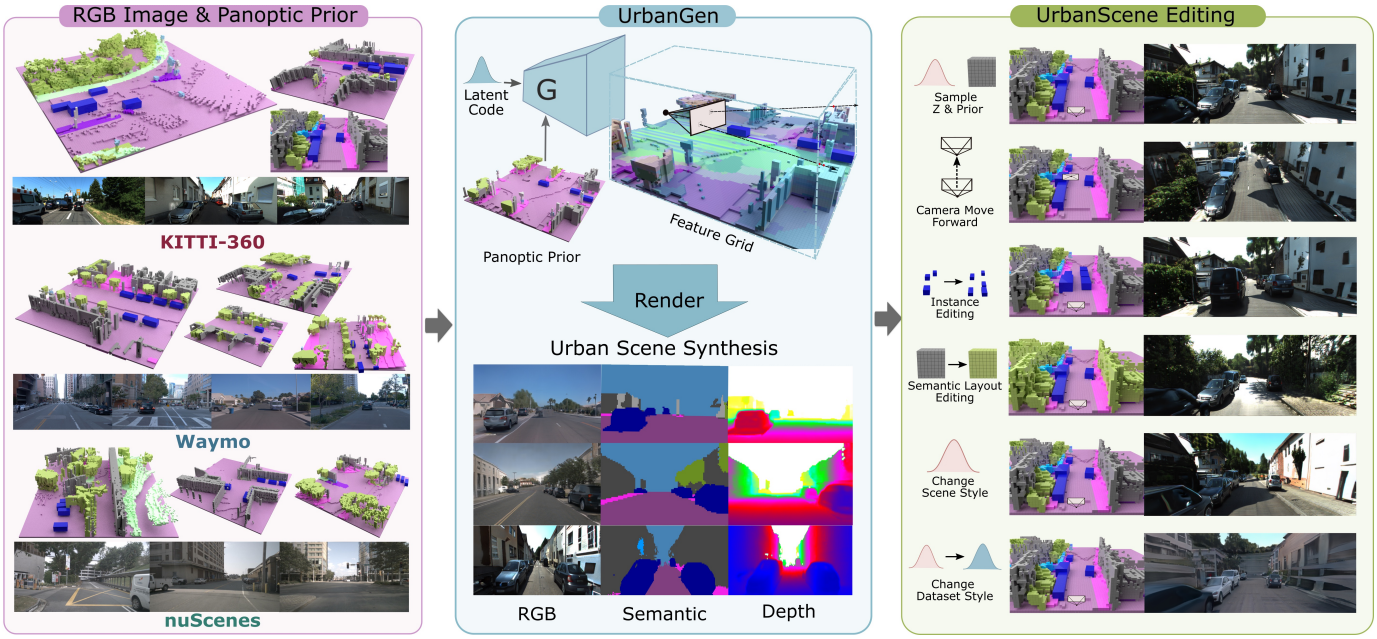


Fig. 1: **UrbanGen Overview.** In UrbanGen, we introduce a 3D generative model for urban scene synthesis. Specifically, we first curate a unified urban dataset from multiple driving datasets, which includes RGB images paired with their panoptic prior, i.e., a combination of scene semantics and 3D bounding box layouts. We then train UrbanGen, a generative neural radiance field that is conditioned on a latent code, dataset style (such as KITTI, Waymo, or nuScenes), and the panoptic prior. UrbanGen can synthesize high-fidelity, diverse urban scenes with multiple modalities (RGB, depth, and semantic) through volume rendering. Finally, thanks to the design of our pipeline, UrbanGen allows for multiple controls over synthesis results, including control over camera pose, instance editing, stuff semantic editing, global appearance editing, and dataset style editing.

dition a stuff generator, whereas the bounding boxes place generated objects into the scene. Specifically, we propose a semantic voxel-conditioned stuff generator via spatially adaptive modulation, effectively injecting the semantic and geometry information provided by the prior. For objects, we follow existing work [45], [67] to generate objects in canonical space, leveraging the cross-instance alignment enabled by the bounding box layout. We further model the sky and far regions using a 3D background generator. With all three generators, we render a composited feature map via volume rendering and upsample it to the target image using a neural renderer. The model is supervised by adversarial losses at both the image and object levels. Furthermore, we propose a semantic alignment loss to encourage the semantic meaning of the rendered image to align with the semantic voxel grid, thereby enabling semantic editing. A geometric loss based on monocular depth prediction is further proposed to improve background geometry not covered by the semantic voxel grid. By constructing unified panoptic priors on representative autonomous driving datasets, including KITTI-360 [34], Waymo [57], and nuScenes [8], we train a city-style aware model conditioned on corresponding style label to synthesize diverse styles across different datasets. Using our proposed training strategy of firstly training on the unified dataset and fine-tuning on each individual dataset, our model achieves high-fidelity synthesis. Training on the unified dataset further enables style control across datasets, such as generating a scene with Waymo's layout in KITTI-360's style. We summarize our contributions as follows:

- We study the novel yet challenging task of urban radiance field generation with rich control in terms of camera pose, object, stuff, and city style.
- We leverage a coarse 3D panoptic-prior to address this challenging task and design compositional generative radiance fields with semantic and geometric losses to leverage the prior information effectively. This additionally enables rendering semantic and depth maps along with RGB images.
- We have collected a unified urban dataset containing images paired with panoptic priors across multiple domains, including KITTI-360, nuScenes, and Waymo. This unification yields higher fidelity than training solely on individual datasets and additionally enables city style transfer.
- Through extensive qualitative and quantitative experiments, we demonstrate that our method outperforms existing state-of-the-art generative methods in terms of visual quality and controllability.

This journal paper extends our conference paper UrbanGIRAFFE [71] published at ICCV 2023 in the following ways: i) We identify that the reconstruction loss used in UrbanGIRAFFE harms image fidelity, even though it helps maintain semantic alignment between rendered images and the panoptic prior. In UrbanGen, we introduce a semantic alignment loss to achieve a better trade-off between image fidelity and semantic alignment, enabling semantic control without sacrificing rendering quality. ii) We use a more advanced background generator to learn the geometry and

appearance jointly, whereas our conference version fixes the background geometry as an infinitely far-away dome. Furthermore, we introduce a monocular depth predictor-guided geometry loss to enhance geometry prediction. This improves the synthesis quality of fine details and far regions. iii) We collect a unified urban dataset with diverse styles. Paired with the newly proposed, city-style aware design of UrbanGen and the training strategy, we successfully extend our method from applying the KITTI-360 dataset to a variety of popular autonomous driving datasets, which additionally enables city-style control.

2 Related Work

3D-Aware Image Synthesis: In this work, we focus on 3D-aware image synthesis learning from 2D supervision. While early works learn to generate 3D voxel grids [41], [23], recent methods achieve high-fidelity 3D-aware image synthesis leveraging neural radiance fields as the underlying 3D representation [52], [10], [11], [53], [15], [68], [21], [49]. Empowered by 3D-aware generative models, many promising applications have been demonstrated, including semantic editing [56], [55], relighting [58], [31], single-view reconstruction [9], [40] and articulated human generation [74], [46], [4], [25]. However, all aforementioned methods focus on object-centric scenes and assume that the objects can be aligned to a canonical space. Thus, it is non-trivial to extend these methods to complex, unaligned urban scenes.

Scene Level 3D Generative Model: One line of methods purely focuses on scene-level geometry generation [30], [37], [64], i.e., generating semantic voxel grids or signed distance fields without appearance. In contrast, we are interested in learning scene-level generative radiance fields capable of rendering photorealistic images. In this direction, GSN [16] and GAUDI [3] propose to generate unbounded indoor scenes. However, both methods ignore the compositionality of the scene, which makes it harder to achieve high visual fidelity and does not support editing of the scene content. More related to us, a few methods exploit the compositionality of 3D scenes to generate scenes containing multiple objects [33], [67], [42], [45], [69]. GIRAFFE [45] and DiscoScene [67] focus solely on the compositionality of foreground objects, thereby being unable to model complex background geometry in urban scenes. Despite achieving promising controllability of foreground objects, they do not support camera control or editing of urban scene elements. Another group of works [2], [76], [29] opts to use bird's-eye view (BEV) semantic images as their condition. CC3D [2] introduces a conditional generative model synthesizing complex 3D scenes based on 2D BEV semantic scene layouts. BerfScene [76] presents an efficient 3D representation that integrates an equivariant radiance field guided by a BEV map. The equivariance property enables it to generate infinite-scale 3D scenes. However, they overlook the compositional nature of complex scenes during modeling, which limits their controllability, such as translating objects and editing local styles.

Several other impressive works explore diverse aspects of scene-level 3D generation [22], [35], [48], [13], [65], [66]. GANCraft [22] is capable of generating photorealistic images of large 3D semantic block worlds. InfiniCity [35] learns to generate large-scale 3D urban scenes relying on 3D CAD datasets. However, both methods are based on test-time optimization, which necessitates a lengthy optimization period for generating a new scene. SceneDreamer [13], CityDreamer [65], and CityGaussian [66] can respectively generate high-fidelity infinite natural and city landscapes in a feed-forward manner. However, their methods are primarily tailored to aerial or distant landscape data. Thus, it is non-trivial to extend them for generating intrinsic 3D scenes with fine-grained details, such as urban scenes for autonomous driving applications.

2D Driving Video Synthesis: Due to the rapid progress of video generation models like SVD [6] and SORA [7], there has been a series of research exploring the use of 2D generative models for urban video synthesis. MagicDrive [19] and MagicDrive3D [18] propose a method for generating videos of street scenes using BEV semantic maps and control. DriveDreamer [63], [77] also uses HD maps and 3D bounding boxes for controllable driving video generation.

Unlike MagicDrive and DriveDreamer, which focus on specific driving datasets such as nuScenes [8], recent efforts by GenAD [70] and Vista [20] have utilized the large-scale 1700-hour driving dataset OpenDV [70] from the internet to create a generalizable driving world model with high fidelity and versatile controllability. However, due to the lack of 3D representations, video-based methods usually fall short in 3D consistency, and cannot yet perform tasks like fine-grained semantic and geometric editing. Our proposed method instead gains 3D consistency by design and enables versatile controllability.

3 Method

In UrbanGen, our goal is to build compositional generative radiance fields of urban scenes with control over camera pose and scene contents. To address this challenging task, we decompose the urban scene into three main components, including stuff, countable objects, and background regions, see Fig. 2 for an overview. We assume prior distributions are provided for both stuff and objects to disentangle the complex urban scenes. In this section, we first introduce the prior distributions of stuff and objects, respectively. Next, we introduce our compositional generator and discriminator for urban scene generation. Finally, we describe the loss functions, sampling strategy, training strategy, and implementation details.

3.1 Panoptic Prior

We assume a prior distribution of the scene layout is given to form a compositional generative model, which we refer to as "panoptic prior". The panoptic prior briefly describes the spatial distributions of countable objects and uncountable stuff within a certain region. Let $\mathbf{V}, \mathbf{O} \sim p_{\mathcal{V},\mathcal{O}}$ denote a stuff layout $\mathbf{V}$ and an object layout $\mathbf{O}$ sampled from the joint distribution $p_{\mathcal{V},\mathcal{O}}$. We now elaborate on the layout representation of $\mathbf{O}$ and $\mathbf{V}$, respectively.

Countable Object: Following GIRAFFE, the layout distribution of countable objects (e.g., cars) is represented

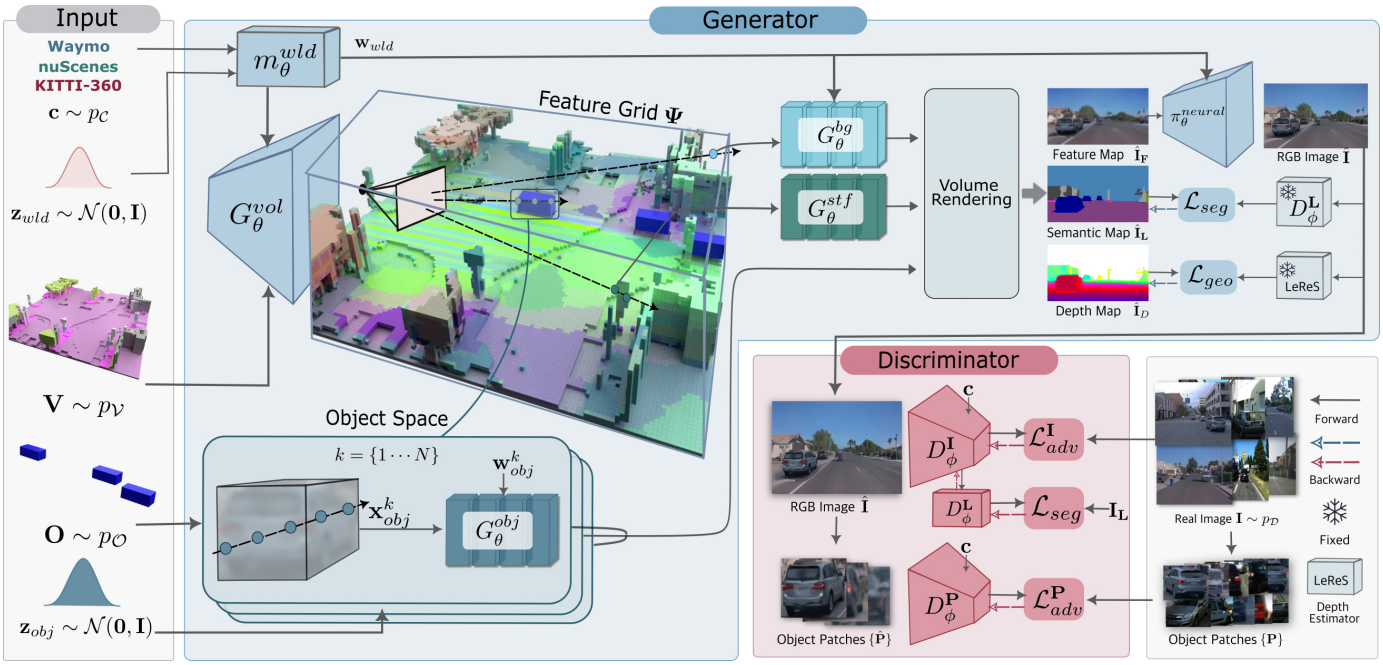


Fig. 2: **Method Overview.** We utilize panoptic prior in the form of semantic voxel grids and instance object layouts to construct a generative urban radiance field. Our model accepts as input a global noise vector $\mathbf{z}_{wld}$ for the entire scene, K noise vectors $\{\mathbf{z}_{obj}^k\}_{k=1}^K$ for objects, a scene domain class $\mathbf{c} \sim p_{\mathcal{C}}$, and a sampled panoptic prior $\mathbf{V}, \mathbf{O} \sim p_{\mathcal{V},\mathcal{O}}$. We decompose the scene into background, stuff, and objects. The stuff generator is conditioned on the semantic voxel grid $\mathbf{V}$ to maintain its semantic and geometric information. Objects are generated in the canonical object coordinate system guided by $\mathbf{O}$. Combined with the background generator, a feature map $\hat{\mathbf{I}}_\mathbf{F}$, depth map $\hat{\mathbf{I}}_D$, and semantic map $\hat{\mathbf{I}}_\mathbf{L}$ are obtained through volume rendering. We further employ neural rendering to produce the RGB image $\hat{\mathbf{I}}$ and object patches $\hat{\mathbf{P}}_k$. The entire model is optimized jointly with adversarial losses $\mathcal{L}_{adv}^\mathbf{I}$ and $\mathcal{L}_{adv}^\mathbf{P}$ applied to the full image and object patches, respectively, as well as a geometry loss $\mathcal{L}_{geo}$ for improved underlying geometry, and a semantic loss $\mathcal{L}_{seg}$ for alignment between the rendered appearance and semantic.

in the form of a set of 3D bounding boxes. A sample $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..\mathbf{o}_K\}$ depicts a joint distribution of K objects in one scene, where K may vary for different scenes. Here, each object $\mathbf{o}$ is represented by a 3D bounding box parameterized by its rotation $\mathbf{R} \in SO(3)$, translation $\mathbf{t} \in \mathbb{R}^3$, and size $\mathbf{s} \in \mathbb{R}^3$:

$$\mathbf{o}_k = \{\mathbf{R}_k, \mathbf{t}_k, \mathbf{s}_k\}$$

In this work, we leverage bounding boxes released by publicly available datasets [34], [8], [1] to form the distribution $p_{\mathcal{O}}$. This distribution can also be obtained from real-world images, e.g., by applying a 3D object detection method.

Uncountable Stuff: Unlike countable objects, there are many indispensable entities that are either uncountable (e.g., road and terrain) or sometimes too cluttered to be separated (e.g., trees). To address this problem, we represent uncountable stuff in the form of semantic voxel grids $\mathbf{V} \in \mathbb{R}^{H_v \times W_v \times D_v \times L}$, where each voxel stores a one-hot semantic label of length L.

3.2 Compositional Urban Scene Generator

Our generator represents the urban scene as a compositional neural radiance field, decomposing the scene into objects, nearby stuff, and far background regions. While similar ideas have been explored in previous works [45], [67], a distinctive aspect of our approach is the modeling of nearby stuff regions through a stuff generator conditioned on a semantic voxel grid, complemented by a background generator for modeling sky and distant areas. Both stuff and background generators are linked by a shared global latent code $\mathbf{z}_{wld} \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ to maintain style coherence throughout the scene. The individual objects are assigned with unique latent codes $\mathbf{z}_{obj} = \{\mathbf{z}_{obj}^k \in \mathcal{N}(\mathbf{0}, \mathbf{I})\}_{k=1}^K$ to ensure a rich diversity in their shapes and appearances. In addition, we also sample a domain label $\mathbf{c}$ corresponding to the dataset domain, such as KITTI-360, Waymo, or nuScenes, and use it as a condition for the generator to assist the model in learning diverse style generation and interpolation using a joint model. We now delve into the details of each part of the generator.

Domain-Aware Mapping Networks: In UrbanGen, we follow [27] to use mapping networks for controlling the generators via latent codes. In addition to taking the random latent codes $\mathbf{z}_{wld}$ or $\mathbf{z}_{obj}$ as input, the mapping networks additionally take the domain label $\mathbf{c}$ as input. This allows for disentangling appearance and dataset domain, thereby guiding the generators to generate high-fidelity images of the given domain more faithfully.

Specifically, the scene mapping network m_θ^{wld} takes as input the global latent code z_{wld} and the domain class label $\mathbf{c}$, and maps them to $\mathbf{w}_{wld}$ which are later used to modulate

Fig. 3: **Scene Domain Editing.** Thanks to training on a unified urban dataset and a domain-aware design of the pipeline, UrbanGen can perform scene domain editing. Specifically, UrbanGen can faithfully render any panoptic prior sampled from one domain into another domain e.g. **Waymo to KITTI**.

the stuff and the background generators:

$$m_\theta^{wld} : (\mathbf{c}, \mathbf{z}_{wld}^k) \mapsto \mathbf{w}_{wld} \tag{1}$$

Similarly, the object mapping network m_θ^{obj} maps the object latent code z_{obj}^k, object scale $\mathbf{s}_k$, and the class label $\mathbf{c}$ to $\mathbf{w}_{obj}$.

$$m_\theta^{obj} : (\mathbf{c}, \mathbf{z}_{obj}^k, \mathbf{s}_{obj}^k) \mapsto \mathbf{w}_{obj}^k \tag{2}$$

Stuff Generator: Our stuff generator generates feature fields for the uncountable stuff condition on the semantic voxel grid $\mathbf{V}$. Inspired by 2D semantic image synthesis [47], [50], we use the semantic voxel grid to modulate the stuff generation. More specifically, our stuff generator consists of a *feature grid generator* G_θ^{vol} and a *MLP head* G_θ^{stf}. The feature grid generator first maps the latent vector $\mathbf{w}_{wld}$ to a feature grid $\mathbf{\Psi} \in \mathbb{R}^{H_v \times W_v \times D_v \times M_v}$ conditioned on the semantic voxel grid $\mathbf{V} \in \mathbb{R}^{H_v \times W_v \times D_v \times L}$:

$$G_\theta^{vol} : (\mathbf{w}_{wld}, \mathbf{V}) \mapsto \mathbf{\Psi} \tag{3}$$

In practice, G_θ^{vol} is a 3D convolutional neural network. The semantic condition $\mathbf{V}$ and the interminate latent codes $\mathbf{w}_{wld}$ are both injected at multiple resolutions using spatially-adaptive normalization, see Fig. 4 as an illustration.

Given a 3D point $\mathbf{x}_{wld}$ within the bounding box $B_\mathbf{V}$ associated with the semantic voxel grid $\mathbf{V}$, we trilinearly interpolate a feature vector $\Psi(\mathbf{x}_{wld}) \in \mathbb{R}^{M_v}$. Next, we map $\mathbf{x}_{wld}$ and $\Psi(\mathbf{x}_{wld})$ to the final stuff feature $\mathbf{f}_{stf} \in \mathbb{R}^{M_f}$ and density σ_{stf} using the MLP head:

$$G_\theta^{stf} : (\Psi(\mathbf{x}_{wld}), \gamma(\mathbf{x}_{wld})) \mapsto (\mathbf{f}_{stf}, \sigma_{stf}) \quad \mathbf{x}_{wld} \in B_\mathbf{V} \tag{4}$$

where $\gamma(\cdot)$ denotes positional encoding. We do not take viewing direction as the input of object and stuff generator, as the neural renderer can learn view dependency effects.

In addition, we can query the semantic label $\mathbf{l} \in \mathbb{R}^L$ of each 3D point $\mathbf{x}_{wld}$ through nearest interpolating the semantic voxel grid $\mathbf{V}$. This allows for obtaining semantic maps via volumetric rendering.

Background Generator: Given the unbounded nature of urban scenes, the stuff generator falls short in modeling distant regions like the sky and skyscrapers, which lie outside of the stuff bounding box $B_\mathbf{V}$. To tackle this, we introduce a background generator to model far regions and sky. Formally, the background generator maps a background point $\mathbf{x}_{wld}$ sampled outside of the stuff bounding box $B_\mathbf{V}$ to a feature vector $\mathbf{f}_{bg} \in \mathbb{R}^{M_f}$, and a density σ_{bg}:

$$G_\theta^{bg} : (\mathbf{w}_{wld}, \mathbf{x}_{wld}) \mapsto (\mathbf{f}_{bg}, \sigma_{bg}) \quad \mathbf{x}_{wld} \notin B_\mathbf{V} \tag{5}$$

Inspired by NeRF++ [75], we utilize inverted sphere parametrization for reparameterizing 3D points in the background regions. It is noteworthy that, in contrast to our conference version [71] which depicts the background as an infinitely distant dome, here we learn the geometry of the background to enhance the photo-realism when moving the camera around. Learning the correct geometry for the background is highly challenging, as there is no geometric guidance as in the stuff generator. Hence, we further propose to incorporate a monocular depth supervision method to enhance the geometry quality of the background regions, as discussed in Section 3.4.

Object Generator: For objects, we follow existing compositional methods to generate each object k in a normalized object coordinate space [33], [45], [67]:

$$G_\theta^{obj} : (\gamma(\mathbf{x}_{obj}^k), \gamma(\mathbf{s}_k), \mathbf{w}_{obj}^k) \mapsto (\mathbf{f}_{obj}^k, \sigma_{obj}^k) \tag{6}$$

where G_θ^{obj} denotes the object generator that maps a 3D point $\mathbf{x}_{obj}^k$, object's size $\mathbf{s}_k$, and a object latent vector $\mathbf{w}_{obj}^k$ to a feature vector $\mathbf{f}_{obj}^k \in \mathbb{R}^{M_f}$ and density σ_{obj}^k. Both $\mathbf{x}_{obj}^k$ and $\mathbf{s}_k$ are encoded by positional encoding $\gamma(\cdot)$. Here, $\mathbf{x}_{obj}^k$ denotes a 3D point in the kth normalized object coordinate, which is transformed to the world coordinate given the object transformation $\{\mathbf{R}, \mathbf{t}, \mathbf{s}\}$:

$$\mathbf{x}_{wld} = \mathbf{R}(\mathbf{s} \odot \mathbf{x}_{obj}^k) + \mathbf{t} \tag{7}$$

Generating objects in this canonical space enables information sharing across different objects, thus allowing for learning a complete shape from many single-view object images. With the learned complete shape, we can control the rotation, translation, and appearance of each individual object.

Note that the scale $\mathbf{s}_k$ is also taken as input to the object generator, as the shape and appearance of vehicles are closely related to their size, e.g., a sedan and a van have notable differences. By adopting this scale-aware approach, the generator is able to capture the scale-based shape and appearance bias, allowing the generated content to more closely align with the real-world data distribution.

Compositional Volume Rendering: We accumulate feature vectors of objects, stuff, and background on each ray via compositional volume rendering. We first sample points from the object, stuff, and background regions independently (the sampling strategy will be elaborated in Section 3.5). Next, we sort all points wrt. their distances to the camera center and accumulate their feature vectors via volume rendering.

Formally, let $\{\mathbf{x}_i\}_{i=1}^M$ denote M sorted points on a ray, compositing of $\mathbf{x}_{obj}^k$ sampled for the object generators (transformed to the world coordinate system via Eq. 7), $\mathbf{x}_{wld} \in B_\mathbf{V}$ for the stuff generator, and $\mathbf{x}_{wld} \notin B_\mathbf{V}$

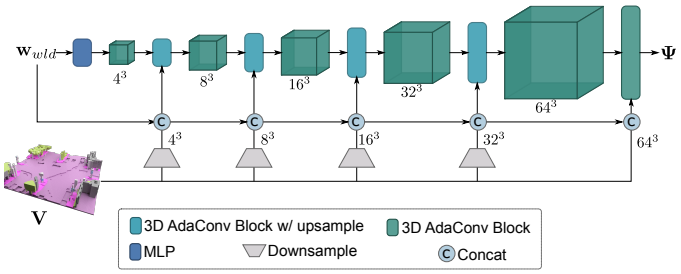


Fig. 4: **Feature Grid Generator** G_θ^{vol} as a part of the stuff generator. We adopt spatially adaptive normalization to inject the semantic condition $\mathbf{V}$ and the latent vector $\mathbf{w}_{wld}$ at multiple resolutions.

for the background generator. For each $\mathbf{x}_i$, we obtain the corresponding feature vector $\mathbf{f}_i$, depth d_i, semantic vector $\mathbf{l}_i$ (points from object and background filed are assigned with semantic 'sky' and 'object'), and density σ_i at $\mathbf{x}_i$. The volume rendering is

$$\pi^{vol} : \{\sigma_i, \mathbf{f}_i, d_i, \mathbf{l}_i\}_{i=1}^{M} \mapsto \{\mathbf{F}, D, \mathbf{L}\} \tag{8}$$

Specifically, the 2D feature $\mathbf{F}$, depth value D and semantic label $\mathbf{L}$ of a ray are obtained via numerical integration as

$$\mathbf{F} = \sum_{i=1}^{N} T_i \alpha_i \mathbf{f}_i \quad D = \sum_{i=1}^{N} T_i \alpha_i d_i \quad \mathbf{L} = \sum_{i=1}^{N} T_i \alpha_i \mathbf{l}_i \tag{9}$$

$$\alpha_i = 1 - e^{(-\sigma_i \delta_i)} \qquad T_i = \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{10}$$

where T_i and α_i denote transmittance and alpha value of a sample point $\mathbf{x}_i$. By compositing all rays, we obtain the rendered feature map $\hat{\mathbf{I}}_\mathbf{F} \in \mathbb{R}^{H_f \times W_f \times M_f}$, depth map $\hat{\mathbf{I}}_D \in \mathbb{R}^{H_f \times W_f}$, and semantic map $\hat{\mathbf{I}}_\mathbf{L} \in \mathbb{R}^{H_f \times W_f \times L}$.

2D Neural Rendering: Following [45], we adopt a neural renderer to transform the rendered feature map to an output RGB image at the target resolution. This allows us to scale to a higher resolution without extensive computation burden. More specifically, our 2D neural renderer π_θ^{neural} maps the feature image $\hat{\mathbf{I}}_\mathbf{F} \in \mathbb{R}^{H_f \times W_f \times M_f}$ and the latent vector $\mathbf{w}_{wld}$ to the RGB image $\hat{\mathbf{I}} \in \mathbb{R}^{H \times W \times 3}$ at the target resolution. Here, $\mathbf{w}_{wld}$ is adopted to enable content-aware upsampling.

$$\pi_\theta^{neural} : (\hat{\mathbf{I}}_\mathbf{F}, \mathbf{w}_{wld}) \mapsto \hat{\mathbf{I}} \tag{11}$$

Note that we directly perform nearest neighbor interpolation to upsample the depth and semantic maps to the final resolution, yielding $\hat{\mathbf{I}}_D \in \mathbb{R}^{H \times W}$ and $\hat{\mathbf{I}}_\mathbf{L} \in \mathbb{R}^{H \times W \times L}$. Here, we reuse the notations of the low-resolution outputs for brevity.

3.3 Compositional Scene Discriminator

UrbanGen, as a conditional generative model, aims not only to produce photorealistic outputs but also to ensure these outputs are aligned with conditioned panoptic priors. This entails aligning rendered appearances with semantic voxel grids and disentangling objects and stuff. While our model is conditioned on the panoptic prior, the semantic information may not be preserved during the adversarial training, as the adversarial loss only distinguishes whether an image is real or not. Therefore, we propose a semantic-aware scene discriminator and integrate an object discriminator to enforce the alignment between the panoptic prior and the synthesized image. Below, we describe these two discriminators in more detail.

Semantic-Aware Scene Discriminator: We propose a semantic-aware scene discriminator and apply it to the full image. To ensure that the generator's outputs are semantically aligned with the input panoptic priors, a discriminator capable of capturing semantic information is essential. Therefore, in addition to predicting whether the input RGB image is real or fake, our scene discriminator further predicts the semantic labels from the input. Specifically, we add a separate branch $D_\phi^\mathbf{L}$ to the scene discriminator for semantic segmentation. Incorporating the semantic prediction allows us to encourage the rendered RGB image to be consistent with the provided panoptic prior, see Section 3.4 for details.

Object Discriminator: Vehicles play a very important role in urban scenes. Generating photorealistic foreground objects, i.e., vehicles, decomposed from the stuff and background is crucial, enabling various downstream tasks like creating diverse traffic scenarios. However, objects often occupy only a small portion of the image, where the scene discriminator applied to the full image provides insufficient supervision for individual objects. Observing this, we introduce an additional object discriminator that takes object patches as input, providing supervision focused at the object level. Specifically, we additionally generate a set of object patches $\{\hat{\mathbf{P}}\}$ for training, as illustrated in Fig. 2.

3.4 Loss Functions

We train the entire model end-to-end using adversarial training. In addition, we propose to utilize a semantic loss to generate 3D scenes semantically aligned with the panoptic prior and a geometry loss to improve the geometry.

Adversarial Loss: We provide adversarial losses to the full image and the objects, separately. Let G_θ denote the full conditional generator that maps the noise vectors and the panoptic-prior to RGB, depth, and semantic label:

$$G_\theta : (\mathbf{z}_{wld}, \mathbf{z}_{obj}, \mathbf{V}, \mathbf{O}) \mapsto (\hat{\mathbf{I}}, \{\hat{\mathbf{P}}\}, \hat{\mathbf{I}}_D, \hat{\mathbf{I}}_\mathbf{L}) \tag{12}$$

Our scene discriminator $D_\phi^\mathbf{I}$ provides adversarial loss to the full image. Specifically, we apply the non-saturated adversarial loss with R1-regularization [38]:

$$\begin{aligned}\mathcal{L}_{adv}^\mathbf{I} =& \mathbb{E}_{\mathbf{I} \sim p_\mathcal{D}} \left[f(-D_\phi^\mathbf{I}(\mathbf{I})) \; - \; \lambda \|\nabla D_\phi^\mathbf{I}(\mathbf{I})\|^2 \right] + \\ & \mathbb{E}_{\mathbf{z}_{wld}, \mathbf{z}_{obj} \sim \mathcal{N}, \mathbf{V}, \mathbf{O} \sim p_{\mathcal{V},\mathcal{O}}} \left[f(D_\phi^\mathbf{I}(\hat{\mathbf{I}}; G_\theta)) \right]\end{aligned} \tag{13}$$

where $(\hat{\mathbf{I}}; G_\theta)$ indicates that $\hat{\mathbf{I}}$ is parameterized by G_θ.

In addition, we adopt object-level discriminative training by feeding the object patches $\hat{\mathbf{P}}$ to another object discriminator $D_\phi^\mathbf{P}$, leading to the object-level adversarial loss $\mathcal{L}_{adv}^\mathbf{P}$:

$$\begin{aligned}\mathcal{L}_{adv}^\mathbf{P} =& \mathbb{E}_{\mathbf{P} \sim p_\mathcal{D}} \left[f(-D_\phi^\mathbf{P}(\mathbf{P})) \; - \; \lambda \|\nabla D_\phi^\mathbf{P}(\mathbf{P})\|^2 \right] + \\ & \mathbb{E}_{\mathbf{z}_{wld}, \mathbf{z}_{obj} \sim \mathcal{N}, \mathbf{V}, \mathbf{O} \sim p_{\mathcal{V},\mathcal{O}}} \left[f(D_\phi^\mathbf{P}(\hat{\mathbf{P}}; G_\theta)) \right]\end{aligned} \tag{14}$$

Semantic Alignment Loss: The segmentation branch is trained on real image-semantic pairs and further used to guide the generator to synthesize 3D scenes aligned with the semantic voxel grid. Let $D_\phi^{\mathbf{L}}$ denote the semantic segmentation branch of $D_\phi^{\mathbf{I}}$ that produces a L-channel semantic map, indicating the semantic probability for each pixel. The semantic loss is formulated as

$$\begin{aligned}\mathcal{L}_{seg} =& \mathbb{E}_{\mathbf{I}\sim p_{\mathcal{D}}}\left[\mathcal{H}(D_\phi^{\mathbf{L}}(\mathbf{I}), \mathbf{I}_{\mathbf{L}})\right] + \\ & \mathbb{E}_{\mathbf{z}_{wld},\mathbf{z}_{obj}\sim\mathcal{N},\mathbf{V},\mathbf{O}\sim p_{\mathcal{V},\mathcal{O}}}\left[\mathcal{H}(D_\phi^{\mathbf{L}}(\hat{\mathbf{I}}; G_\theta), \hat{\mathbf{I}}_{\mathbf{L}})\right] \end{aligned} \quad (15)$$

where $\mathcal{H}(.,.)$ denotes pixel-wise cross-entropy loss and $(\mathbf{I}, \mathbf{I}_{\mathbf{L}})$ stands for a ground-truth image-semantic pair. The semantic label $\mathbf{I}_{\mathbf{L}}$ of a real image can be either obtained from the dataset or using pre-trained semantic segmentation models, whereas the semantic label $\hat{\mathbf{I}}_{\mathbf{L}}$ of the synthesized image is obtained via volume rendering. Here, the first part of $\mathcal{L}_{seg}$ guides the *discriminator* $D_\phi^{\mathbf{L}}$ to learn semantic segmentation given an input image, whereas the second part guides the *generator* to produce a rendered image $\hat{\mathbf{I}}$ to be semantically aligned with the label $\hat{\mathbf{I}}_{\mathbf{L}}$ obtained from volumetric rendering. This encourages the generated scene to maintain the panoptic prior as the semantic label $\hat{\mathbf{I}}_{\mathbf{L}}$ is directly obtained through querying the semantic voxel grid. Our semantic alignment loss draws inspiration from 2D semantic image synthesis methods, such as OASIS [51]. Unlike these 2D methods, which focus on converting a 2D semantic map into an RGB image where the semantic label $\hat{\mathbf{I}}_{\mathbf{L}}$ is known and precise, we employ a rendered semantic label $\hat{\mathbf{I}}_{\mathbf{L}}$ derived from a coarse 3D panoptic prior.

Geometry Loss: Although the semantic voxel grid $\mathbf{V}$ provides a coarse geometric prior for the scene, relying solely on an adversarial loss on RGB images is still insufficient for learning the finer-grained underlying geometry. Moreover, the background regions have no geometric guidance to produce the correct geometry. Therefore, we propose to leverage pixel-wise depth supervision to encourage our generator to learn more realistic geometry.

Specifically, we utilize an off-the-shelf depth estimation model [72] to obtain pseudo-ground truth depth maps $\hat{\mathbf{I}}_D^r$ for the rendered RGB images $\hat{\mathbf{I}}$. We then apply a mean squared error to supervise the depth maps $\hat{\mathbf{I}}_D$ generated by volume rendering. The geometry loss for the generator can be formulated as follows:

$$\mathcal{L}_{geo} = \mathbb{E}_{\mathbf{z}_{wld},\mathbf{z}_{obj}\sim\mathcal{N},\mathbf{V},\mathbf{O}\sim p_{\mathcal{V},\mathcal{O}}}\left[\|(\hat{\mathbf{I}}_D; G_\theta) - \mathbf{I}_D^r\|_2^2\right]$$

Full Objectives: In summary, the generator G_θ and the discriminators $D_\phi^{\mathbf{I}}$ and $D_\phi^{\mathbf{P}}$ are jointly optimized with

$$\mathcal{L} = \mathcal{L}_{adv}^{\mathbf{I}} + \lambda_{\mathbf{P}}\mathcal{L}_{adv}^{\mathbf{P}} + \lambda_{seg}\mathcal{L}_{seg} + \lambda_{geo}\mathcal{L}_{geo} \quad (16)$$

where $\lambda_{\mathbf{P}}, \lambda_{seg}$ and λ_{geo} are loss weights to balance different terms.

3.5 Sampling Strategy

We use the panoptic prior to guide the sampling of volume rendering, effectively reducing the required number of sampling points and improving rendering efficiency. Additionally, we design a camera pose sampling strategy during training to further improve training efficiency.

Ray-Voxel Intersection Sampling for Stuff: Inspired by existing methods [22], [36], we use the ray-voxel intersection sampling strategy to determine sampling locations for the stuff generator. For each ray, we find the first 4 non-empty voxels that the ray hits and then sample M_{vol} points within each of these voxels. This effectively reduces the number of required sampling points by avoiding sampling in the empty space and occluded regions.

Ray-Box Intersection for Object: For objects, we also leverage the 3D bounding boxes to reduce the number of samples in the empty space. Given a ray, we first calculate the ray-box intersections for each bounding box parameterized by $(\mathbf{R}, \mathbf{t}, \mathbf{s})$. Next, we sample M_{obj} points within each bounding box by uniform sampling between the intersections. We use the stratified sampling strategy following [39], i.e., a random shift is added to the sampled points.

Camera Pose Sampling: During training, we sample from a set of plausible camera poses for each panoptic layout. This ensures that the synthesized radiance field can render photorealistic images across multiple viewpoints. The plausible camera poses are extracted from each dataset using the driving trajectory corresponding to each panoptic layout. However, the unfiltered camera poses can present several challenges, such as exiting the scene box, turning around, or reversing direction. To address these issues, we implement a camera sampling strategy following CC3D [2]. Specifically, we initially filter out camera poses that fall outside the region of the semantic voxel grid. Subsequently, we impose constraints on the camera's facing direction, ensuring it deviates by no more than 45 degrees from the forward direction of the semantic voxel grid. This sampling strategy ensures that the areas of interest are within the modeling scope of the stuff generator, facilitating effective training of the model. We sample 15 poses satisfying the above conditions for each panoptic layout.

3.6 Training Strategy

Existing generative radiance fields typically train one specific model for each dataset. This approach face challenges when one dataset is of small scale and prevents from leveraging the information across diverse datasets. As detailed in Section 4.1, we collect a unified dataset with images and panoptic priors in the same format across multiple datasets, including KITTI-360, Waymo, and nuScenes. This provides an alternative training method, where a unified model can be trained on all datasets jointly. While this unified model achieves reasonable performance and enables city-style transfer, its fidelity on a single dataset may be further enhanced considering the large city style difference across multiple datasets. Therefore, we propose an effective training strategy. Specifically, we first train on the unified dataset of multiple datasets until convergence. This unified model is used for city-style transfer experiments. We then fine-tune this unified model for each dataset to achieve higher image fidelity. We validate the effectiveness of this training strategy in our ablation study.

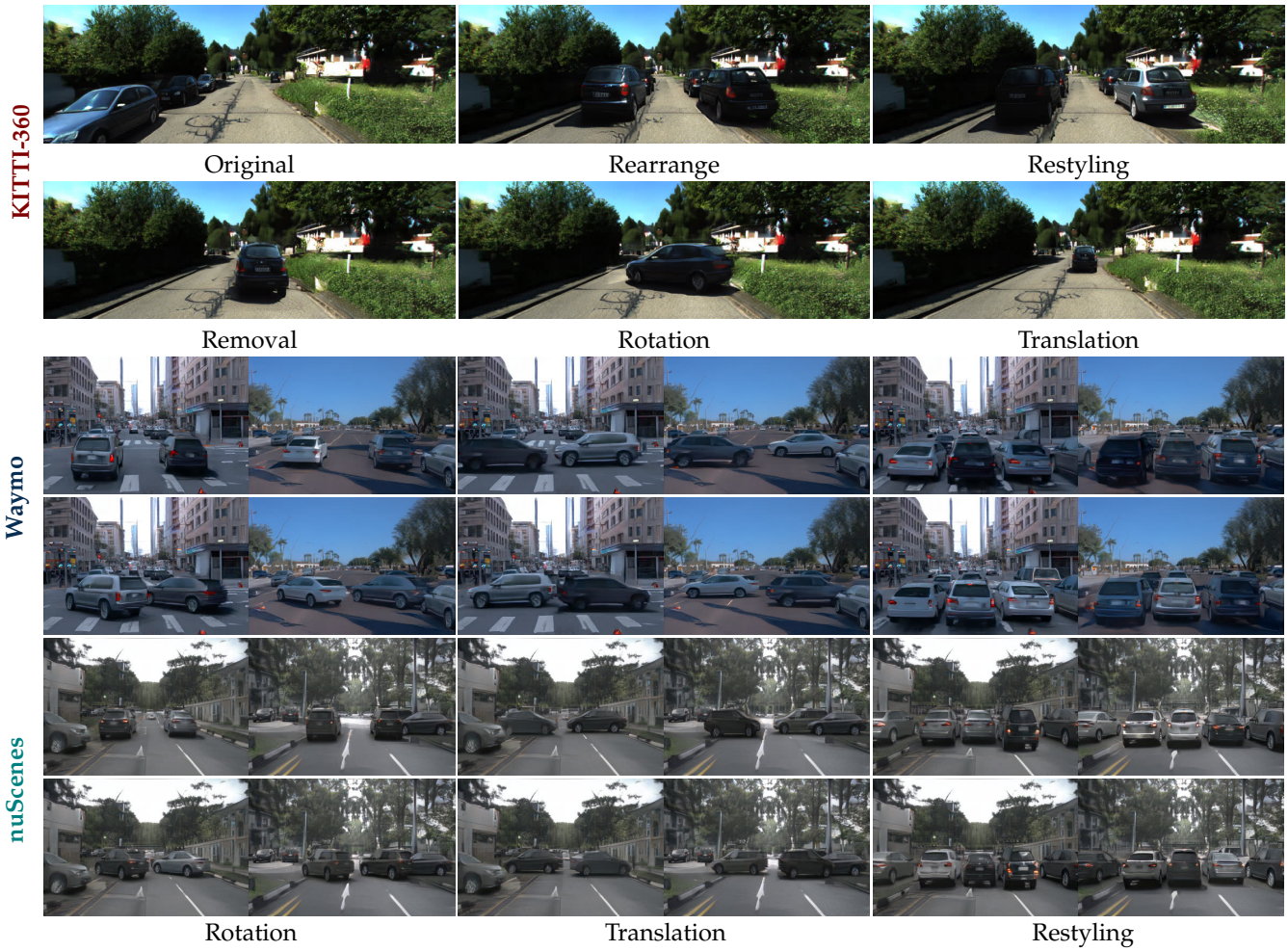


Fig. 5: **Object Editing.** We perform versatile control of scene objects across multiple datasets, including rearrangement, removal, insertion, rotation, translation, and restyling of these objects. With the help of panoptic prior, UrbanGen accurately separates the background from the objects, ensuring that the appearance of other parts of the scene remains consistent during the object editing process.

3.7 Implementation Details

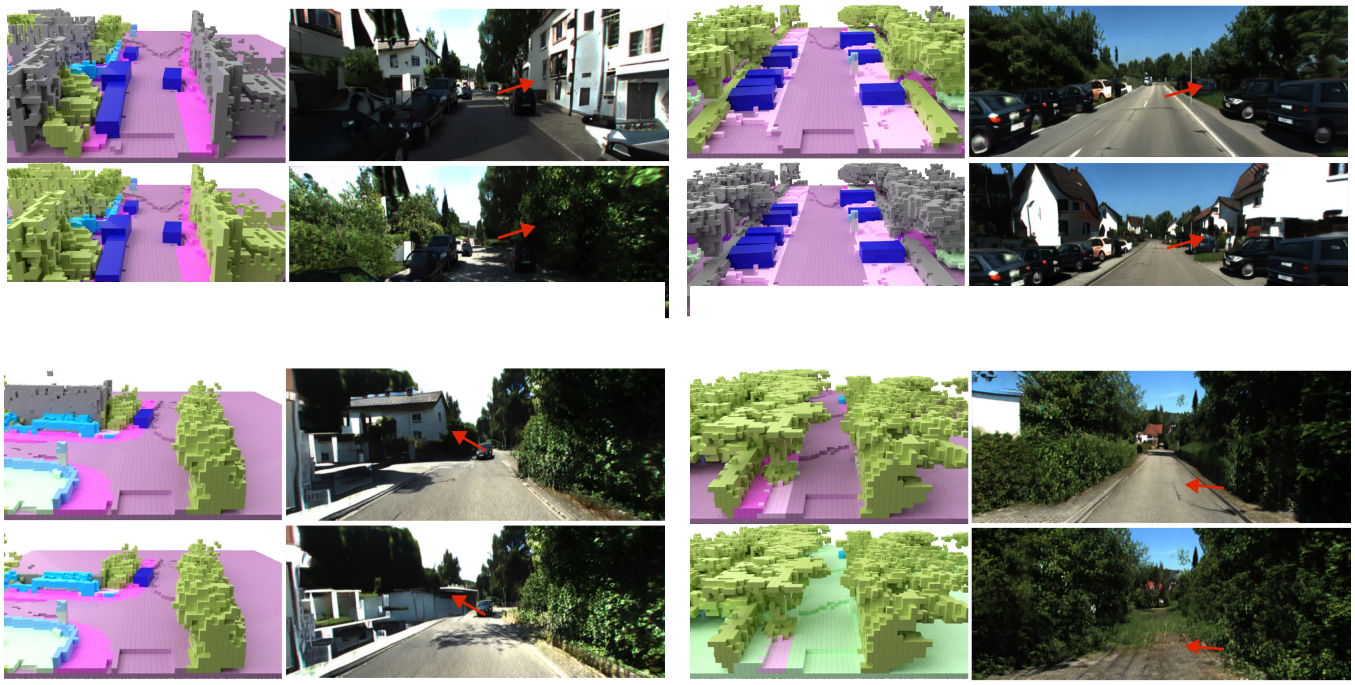
Generator: We use 3D CNNs with 5 spatially-adaptive normalization blocks for the stuff generator G_θ^{vol}. We set $H_v = W_v = 64$ and $D_v = 16$ for all experiments, i.e., the semantic voxel grids are at the resolution of $64 \times 64 \times 16$. We use $M_v = 32$ channels for the feature grid $\mathbf{\Psi}$ to avoid large memory consumption. Regarding mapping networks, both the object mapping module m_{obj}^θ and world mapping module m_{wld}^θ use the same architecture and parameters as the conditional mapping network from EG3D [10]. For object generator G_θ^{obj}, background generator G_θ^{bg} and the MLP head of stuff generator G_θ^{stf}, we use 8, 4, and 4 modulated fully-connected layers with 128 channels, respectively.

Discriminator: Both discriminators $D_\phi^{\mathbf{I}}$ and $D_\phi^{\mathbf{P}}$ share the similar EG3D's dual-discriminator for better multi-view consistency. The full image discriminator $D_\phi^{\mathbf{I}}$ takes as input images at the resolution of 256×256 pixels. We randomly crop or resize real images and fake images to 256×256, each with a probability of 50%. The input resolution of the patch discriminator $D_\phi^{\mathbf{P}}$ is 128×128.

Training: During training, we sample camera poses along plausible driving trajectories given a semantic voxel grid. In terms of the resolution of the rendered feature map $\hat{\mathbf{I}}_{\mathbf{F}}$, we set $H_f = 64$, and W_f is determined by the aspect ratio of different datasets ($W_f = 96$ on nuScenes and Waymo, and $W_f = 192$ on KITTI-360) with $M_f = 32$. The neural renderer then upsamples the feature map to $\hat{\mathbf{I}}$ with a resolution of 256×384 on nuScenes and Waymo, and 256×768 on KITTI-360. Regarding ray marching, we sample $M_{obj} = 12$ points within each object's bounding box and $M_{vol} = 6$ within each voxel. All our models are trained on 8 * A800 GPUs with a batch size of 64. Unless specified, other training and inference hyperparameters are the same as EG3D [10]. Our model is first trained on the Unified Urban Dataset for 128K iterations. To achieve better visual fidelity for specific domains, we further fine-tune the model for each city style with an additional 12.8K iterations on Waymo and 6.4K iterations on nuScenes and KITTI-360.

4 Experiments

In this section, we first compare our method to several 2D and 3D baselines on generation fidelity and geometric quality across multiple real-world urban datasets. Subsequently, we design several controllable urban scene editing



Building Lower Road to Grass

Fig. 6: **Stuff Semantic Editing on KITTI-360** can be efficiently achieved by editing semantic voxel grids. In each group, the first row presents the unedited semantic voxel grid sampled from the dataset along with the synthesized image rendered by UrbenGen. The second row presents an edited semantic voxel grid and its corresponding image.

experiments to evaluate the preferences of our synthesis model with regard to controllability and fidelity. We further conduct ablation studies to better understand the influence of different architectural components.

4.1 Datasets

We conduct experiments on three autonomous driving datasets: KITTI-360, nuScenes, and Waymo. To enable diverse style generation with a single model, we further construct a unified dataset combining all these three datasets.

KITTI-360 [34]: KITTI-360 captures sub-urban regions in Germany, encompassing diverse sensor modalities as well as panoramic viewpoints. We utilize all publicly available sequences from KITTI-360 for training. Specifically, we use all frames captured by the left perspective camera, resulting in a total of approximately 60K training images. We leverage 3D object annotations provided within the dataset and semantic voxel grids from the SSCBench [32] dataset to form panoptic priors. We use the 2D semantic labels provided by this dataset to supervise the semantic alignment loss.

Waymo [1]: The Waymo Open Dataset is collected from various locations across the United States, consisting of 1,000 short sequences for training and validation with each sequence spanning 20 seconds. For UrbanGen, we utilize all front-view images from the training scenes. We manually filtered out scenes captured during nighttime and rainy conditions, resulting in a curated subset of approximately 120K frames, each labeled with instance bounding boxes. Additionally, we acquired the corresponding semantic voxel grids from the Occ3D [61] dataset. The 2D semantic labels are predicted by a pre-trained semantic segmentation model [59].

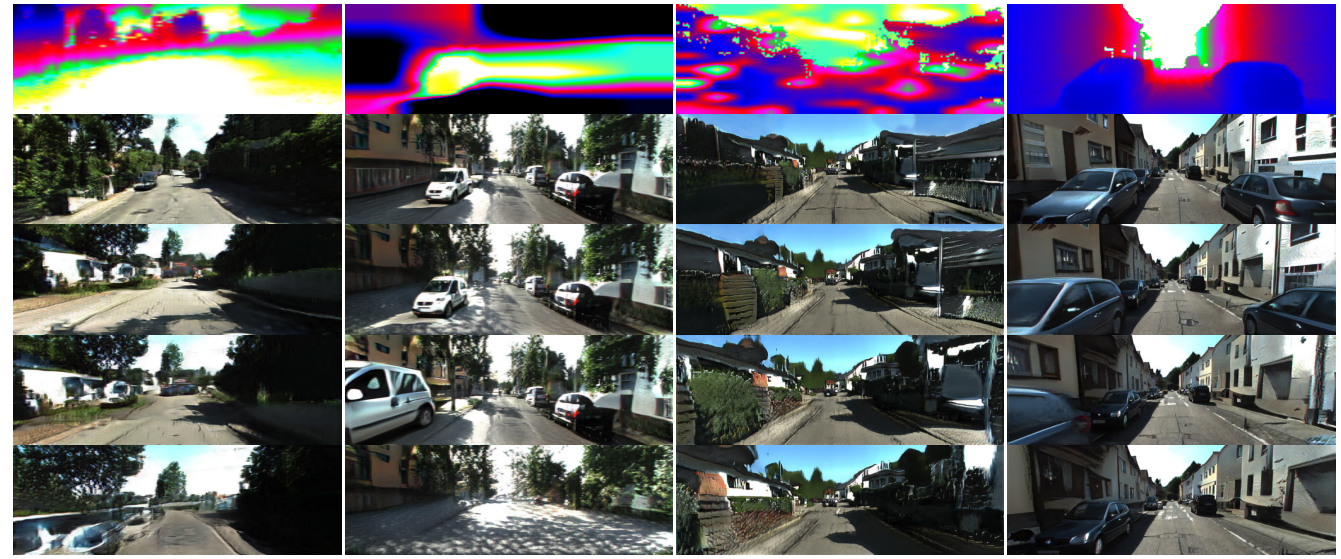
nuScenes [8]: The nuScenes dataset comprises 850 training and validation sequences, each lasting 20 seconds. Given that nuScenes offers annotations for keyframes at a 2Hz frequency, UrbanGen utilizes only the front camera images of these keyframes for training purposes, and excludes scenes captured during rainy and night conditions, culminating in a total of approximately 28K frames. Same as Waymo, the semantic voxel grids are obtained from Occ3D [61] and the 2D semantic labels are predicted by [59].

Unified Urban Dataset: To enable the urban generation of diverse styles, we develop a Unified Urban Dataset by combining all three datasets mentioned above for training. Specifically, we establish a label mapping mechanism to unify the semantic labels of KITTI-360, Waymo, and nuScenes. Based on the label mapping proposed in SSC Bench [32], we map the labels from Waymo, nuScenes, and KITTI-360 to the Cityscapes [14] semantic map. Additionally, we introduce extra semantic labels, such as "traffic cone" and "general object", to accommodate the unique semantics present in each of these datasets. Moreover, we crop and downsample the semantic voxel grids of these datasets to obtain a semantic voxel grid with the same resolution of $64 \times 64 \times 16$, covering 51.2×51.2 square meters in the bird eye view, and 6.4 meters in height. This results in a unified dataset consisting of 200,000 frames.

4.2 Metrics

Our evaluation metrics measure the visual quality of generated content, 3D consistency, and the quality of the underlying geometry.

FID and KID: We use the standard Fréchet Inception Distance (FID)[24] and Kernel Inception Distance (KID)[5] scores to measure image quality. During the evaluation, we



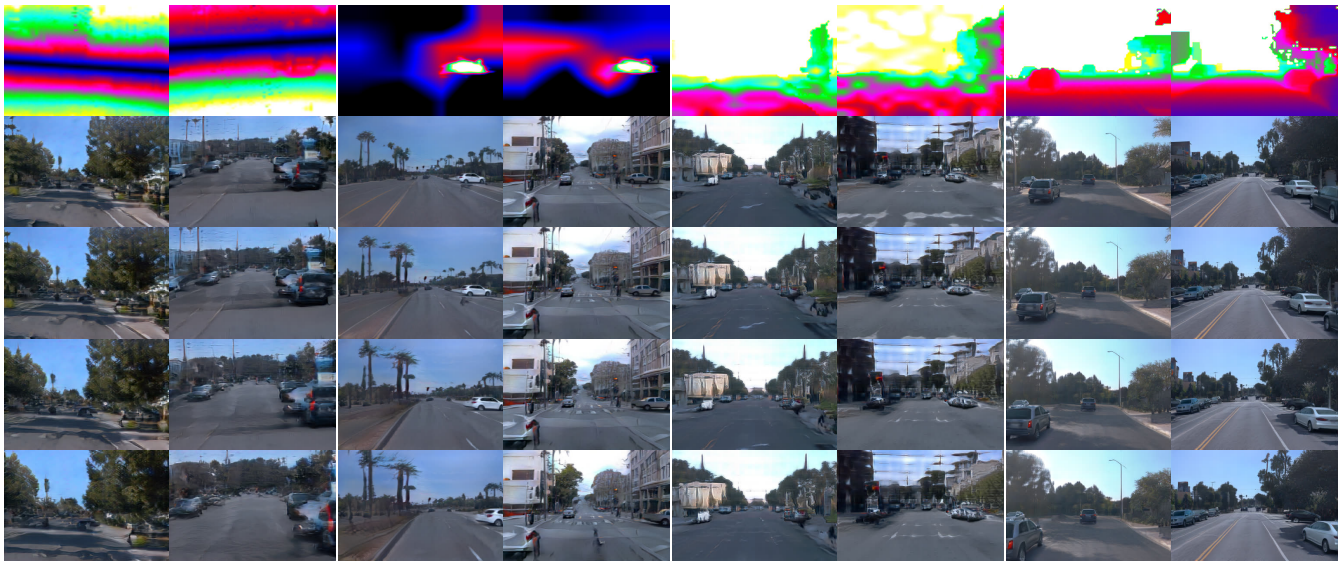


Fig. 7: **Qualitative Comparison on KITTI-360 and Waymo**. The first two rows present RGB images paired with depth, rendered by each method from the default camera pose located at the origin of the scene. In the following rows, the camera moves forward, accumulating a total moving distance of 10 meters.

sample 50k fake examples and all real samples from the dataset to calculate the FID and KID scores.

Depth Error: Following existing 3D-aware GANs [10], [13], we measure geometry quality via a Depth Error (DE) metric. We adopt a pre-trained monocular depth estimation model [73] to generate a pseudo ground truth depth map for every rendered image. We calculate the L2 distance between the pseudo-GT depth maps and the rendered depth maps, both normalized to zero mean and unit variance to mitigate scale discrepancies.

Camera Error: The Camera Error (CE) metric assesses multi-view rendering consistency by quantifying the deviations between inferred and SfM-estimated camera trajectories. In practice, we use the recently released powerful DUST3R [62] as the pose estimator to estimate the camera poses of rendered sequences. Before calculating the Camera Error, we normalize all camera poses. Specifically, we first convert all camera poses to be relative to the first one, so that the first camera pose becomes an identity matrix. Then, we re-scale the translation to ensure that the distance from the first to the farthest camera is set to 1.

4.3 Comparison to the State of the Art

We compare our method against three state-of-the-art models. EG3D [10] introduces a tri-plane representation that significantly enhances the efficiency and rendering quality of 3D GAN. DiscoScene [67] is a generative radiance field that uses 3D bounding boxes as scene layout priors to spatially disentangle the scene into objects and a background. CC3D [2] is closely related to our work, which is a conditional generative model capable of synthesizing complex 3D scenes based on 2D semantic scene layouts. To evaluate the image synthesis fidelity, we also compare our method with StyleGAN2 [28], a renowned 2D image generation method.

Method	KITTI-360				Waymo		NuScenes		Unified	
	FID ↓	KID ↓	DE ↓	CE ↓	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓
StyleGAN2 [28]	7.2	5.1	–	–	**6.1**	**7.2**	**7.6**	9.7	12.1	11.2
EG3D [10]	11.9	22.4	0.612	0.343	32.7	41.7	46.3	53.5	52.6	58.5
DiscoScene [67]	34.7	20.9	0.892	0.647	25.6	19.1	31.2	28.9	61.2	55.4
CC3D [2]	59.6	62.4	0.334	0.095	71.1	69.0	77.2	88.9	82.0	94.5
UrbanGIRAFFE [71]	43.3	40.0	0.291	0.055	–	—	—	—	—	—
Ours	**6.9**	**4.3**	**0.194**	**0.042**	8.2	7.7	21.4	14.7	**11.2**	**9.5**

TABLE 1: **Quantitative Comparison.** We conduct experiments on KITTI-360, Waymo, nuScenes and Unified urban dataset. FID, KID ($\times 10^3$) are reported as the evaluation metrics of image fidelity. CE, DE are reported as the evaluation metrics of geometry quality. Note that we highlight the best results among all methods.

Rendering Fidelity and Geometric Quality: We first conduct comparisons on **KITTI-360**, Waymo, nuScenes, and the Unified Urban Dataset that combines all three datasets. Table 1 and Fig. 7 show that our method greatly outperforms existing state-of-the-art 3D methods regarding FID and KID and is comparable to the 2D baseline. Specifically, EG3D [10] can learn reasonable results from 2D image collections. However, due to the lack of geometry prior, it struggles to learn the underlying geometry of urban scenes, resulting in low FID and high Depth Errors. DiscoScene [67] shows high-quality foreground objects by incorporating object layout prior and learning objects in their canonical space. However, due to its overly simplified modeling method of urban scene structure, it results in poor geometry, which is reflected by the depth error as shown in Fig. 7. CC3D [2], thanks to the scene semantic prior provided through the BEV semantic map, results in lower depth errors but the FID and KID are unsatisfying due to lack of details. Finally, our conference version, UrbanGIRAFFE, outperforms the other baselines in terms of depth error, demonstrating that the panoptic prior plays an important role in easing the task of learning accurate geometry. By replacing the reconstruction loss with a semantic alignment module, UrbanGen significantly enhances image fidelity compared to UrbanGIRAFFE. Additionally, the incorporation of geometry loss further reduces the depth error in UrbanGen. These design improvements make our method comparable to the 2D baseline, StyleGAN2, and even allow it to outperform StyleGAN2 on some datasets. However, a larger performance gap is observed on the nuScenes dataset due to its limited sample size and diversity, which also poses challenges for other 3D generation methods.

It is worth noting that in the unified dataset setting, all 3D baseline methods fail to generalize across multiple datasets, causing both FID and KID to rise significantly. In contrast, due to our method's domain-aware design and the unified panoptic prior, we can easily learn urban scene representations that span multiple domains, outperforming all methods including 2D baselines in this most challenging setting.

Reconstruction Quality: To better compare the multiview consistency between different methods, we reconstruct the point cloud and camera trajectory from the forward-moving image sequences using Dust3R [62]. The reconstruction results are shown in Fig. 8, where we additionally visualize the corresponding conditions of conditional generative methods. Similar to the conclusions above, EG3D's image sequences exhibit severe artifacts in the reconstruction. DiscoScene's results barely recover the camera trajectory and a reasonable scene. In contrast, CC3D can mostly recover the camera trajectory and 3D scene but still lacks detail. In comparison, Dust3R can successfully recover the camera trajectory and detailed scene, such as vehicles and lane markings, from the image sequences rendered by UrbanGen, validating the multiview consistency. Additionally, UrbanGen's reconstructed point cloud maintains a high degree of semantic alignment with the conditioned panoptic prior.

4.4 Controllable Urban Scene Generation

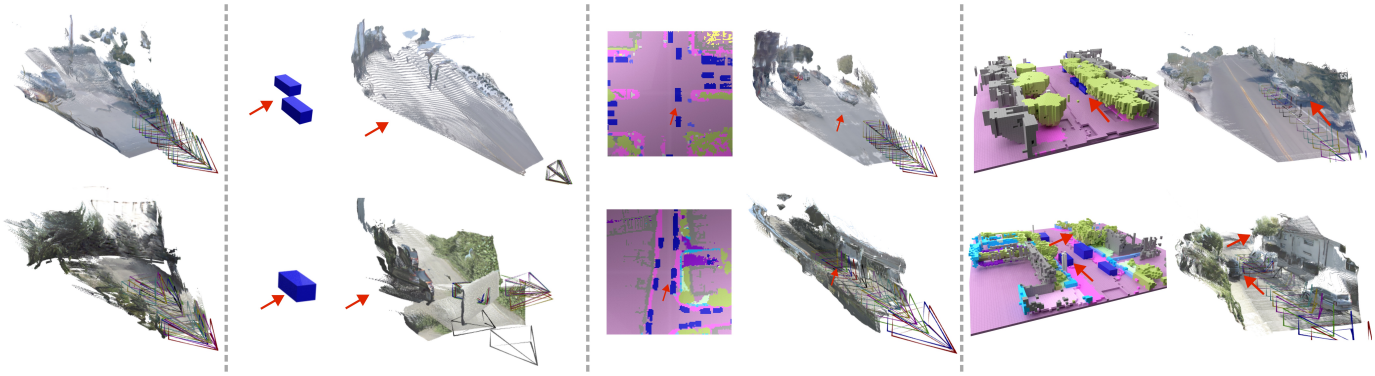
We now demonstrate the diverse controllability of our model in terms of multi-modality rendering, stuff editing, object editing, camera viewpoint control, and scene-style editing.

Multi-modality Image Synthesis: Thanks to the semantic and geometric cues provided by the panoptic prior, UrbanGen is not just capable of generating photorealistic urban scenes but also has a holistic understanding of each scene. This allows for rendering paired depth and semantic images alongside an RGB image, see Fig. 9. While the depth and semantic maps are coarse, they are obtained via volume rendering and, hence are multi-view consistent by design.

Object Editing: Fig. 5 illustrates the object editing capability of UrbanGen. As in GIRAFFE [44] and DiscoScene [68], we can add/delete objects, and control their appearance, rotation, and translation. Results show that editing the objects does not change other parts of the scene, indicating that the panoptic prior allows for disentangling foreground objects from the rest. It is noteworthy that we have achieved promising results across multiple challenging autonomous driving datasets.

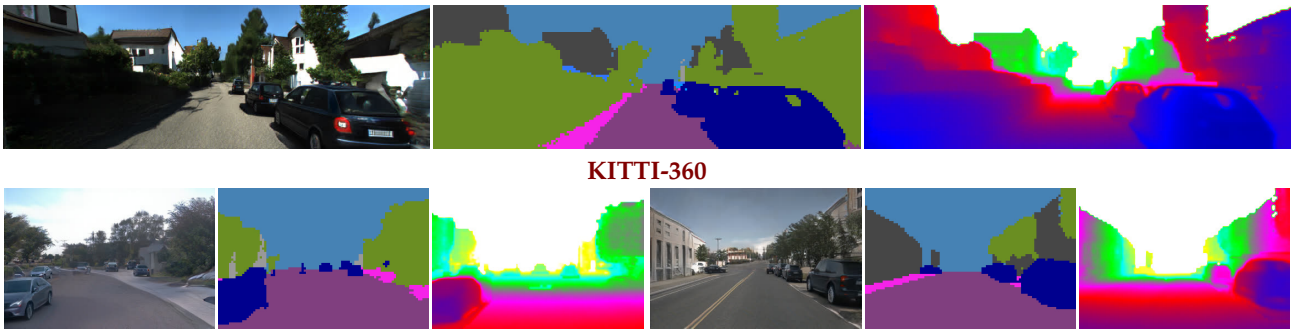
Stuff Editing: Next, we perform experiments in stuff editing. Our semantic-conditioned stuff generator enables fine-grained stuff editing by modifying the conditioning semantic voxel. Such stuff editing can be achieved by performing feed-forward inference conditioned on the modified semantic voxel grid without additional optimization. As shown in Fig. 6. We can manipulate stuff semantics like changing "Tree" to "Building" and vice versa, and changing "Road" to "Grass". It is also possible to edit the occupancy of the voxel grids like lowering the building.

It is worth mentioning that, as shown in the "Building to Tree" example, the shadow of the road also changes



EG3D [10] DiscoScene [67] CC3D [2] UrbanGen (Ours)

Fig. 8: **3D Reconstruction of Synthesis Results.** We first synthesize image sequences for each method by shifting the camera forward 10 meters and then use Dust3R [62] to perform 3D reconstruction. The conditions for each method are provided on the left side of the figure except for EG3D which is unconditional. DiscoScene conditions on object layout, CC3D conditions on bird's-eye-view semantic layout, and UrbanGen conditions on panoptic prior.



KITTI-360

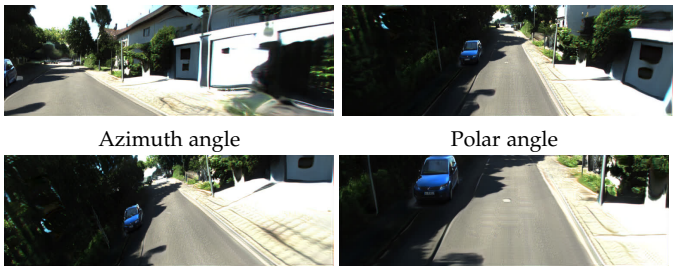
NuScenes **Waymo**

Fig. 9: **Multi-modality Urban Image Synthesis**. UrbanGen allows for simultaneously synthesizing paired RGB, semantic, and depth images of each scene across multiple datasets (**KITTI-360**, **Waymo** and **n uScenes**).

correspondingly after the editing. This suggests that our method not only enables photorealistic and semantically aligned urban scene generation but also learns the implicit relationship between the shadow condition and the semantic layout. Additionally, an interesting failure case occurs when changing "Road" to "Grass", where the close regions remain as road. This may result from the lack of training data representing direct driving on grass.

Camera Control: As shown in Fig. 10, our method also allows for large viewpoint control, including large rotation in azimuth and polar angles as well as in-plane rotation. We can also change the camera's focal length, successfully capturing a photorealistic wide-angle image. The camera controls, which maintain good multi-view consistency, enhance the method's capability to support downstream applications, such as autonomous driving scene simulation.

Style Editing: We now test the ability of scene style editing on our method, which is controlled by latent code $\mathbf{z}_{wld}$. As depicted in Fig. 11, we demonstrate an interpolation between two latent codes to create a smooth transition within the latent space. As seen in the second row of the figure, with the linear alteration of the latent code, a countryside path transitions from a normal sunny day to a bright, sunlit environment, underscoring that the latent code captures style and lighting effects. It is important to note that the semantic layout remains consistent despite the significant



Azimuth angle Polar angle

In-plane rotation Changing FoV

Fig. 10: **Camera Control on KITTI-360 dataset.** We perform versatile camera control, including large viewpoint translations, rotations in azimuth and polar angles, as well as in-plane rotations and changes in the camera's field of view.

transformation in the scene's appearance. This is aligned with the conditioned panoptic prior, showcasing the model can preserve the semantics and geometry prior.

Style Transfer: Fig. 3 evaluate our model's ability to perform scene domain editing, i.e., transferring the style from one dataset domain to another. As mentioned in Fig. 2, different dataset domains correspond to different input labels. By altering the label during inference, we can switch the generated scene's style to any domain, such as transforming a Waymo layout to KITTI-360 or nuScenes style. These domain editing results indicate that UrbanGen learns to

Fig. 11: **Scene Style Editing.** Each row demonstrates global scene style editing by interpolating the latent code $\mathbf{z}_{wld}$, whereas the $\mathbf{z}_{obj}$ for foreground cars remains unchanged. This process facilitates a smooth transition within the latent space, e.g., the shadow changes smoothly on the ground in all three rows, emphasizing the latent code's capability to capture and manipulate stylistic and lighting aspects.

	$\text{FID}_\mathbf{I} \downarrow$	$\text{FID}_\mathbf{P} \downarrow$	DE $\downarrow$	CE $\downarrow$
w/o $\mathcal{L}_{adv}^{\mathbf{P}}$	12.7	78.2	0.263	0.089
w/o G_θ^{obj}	13.1	83.5	0.289	0.101
w $\mathcal{L}_{recon}$	31.1	42.4	0.227	0.067
w/o $\mathcal{L}_{seg}$	**6.6**	28.8	0.199	0.052
w/o $\mathcal{L}_{depth}$	7.3	29.2	0.301	0.069
Full	6.9	**24.8**	**0.194**	**0.042**

TABLE 2: **Ablation Study** conducted on KITTI-360 with different method variations, including omitting the object discriminator (w/o $\mathcal{L}_{adv}^{\mathbf{P}}$), excluding the object generator (w/o G_θ^{obj}), adding reconstruction loss (w $\mathcal{L}_{recon}$), removing depth loss (w/o $\mathcal{L}_{geo}$), and disabling semantic alignment loss (w/o $\mathcal{L}_{seg}$).

disentangle the style of the dataset domain from the panoptic prior, hence being able to generate out-of-distribution scenes, e.g., US layouts with German-style.

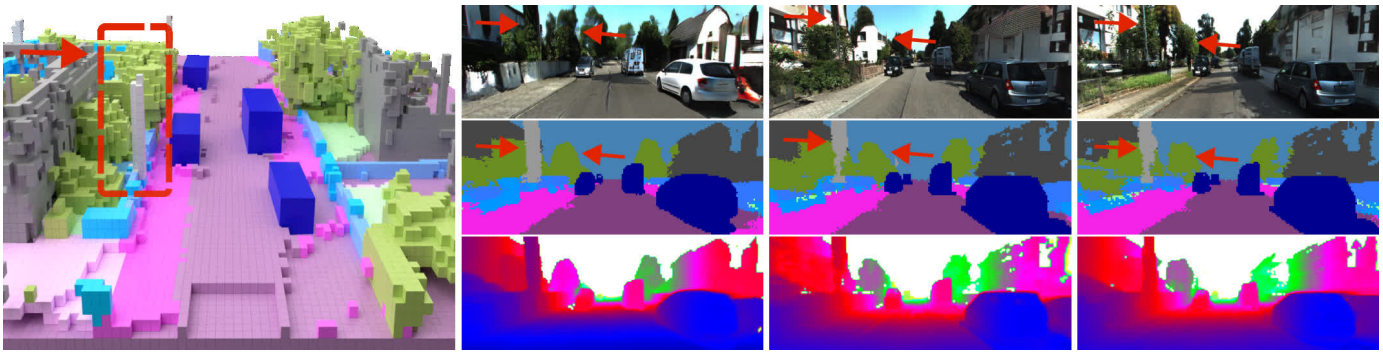
4.5 Ablation Study

To verify our design choices, we conduct ablation studies on the KITTI-360 dataset, and evaluate both image-level and patch-level FID/KID scores in Table 2 where we show how our design choices affect the generative performance of UrbanGen. Specifically, we analyze the following factors: i) the object discriminator; ii) the object generator; iii) the stuff reconstruction loss used in our conference version [71]; iv) the semantic alignment loss; and v) the geometry loss. The ablation study in Table 2 removes one component at a time and evaluates corresponding scene level and patch level FID scores respectively.

Object Discriminator: Firstly, we exclude the adversarial loss $\mathcal{L}_{adv}^{\mathbf{P}}$ applied to object patches and train the object generator solely through the image adversarial loss $\mathcal{L}_{adv}^{\mathbf{I}}$. As shown in Table 2 (w/o $\mathcal{L}_{adv}^{\mathbf{P}}$), removing $\mathcal{L}_{adv}^{\mathbf{P}}$ significantly increases the patch $\text{FID}_\mathbf{P}$ and $\text{KID}_\mathbf{P}$. It is worth noting that $\text{FID}_\mathbf{I}$ is less affected, indicating that in scenes where the proportion of object pixels is not large, the global adversarial training cannot provide enough supervision to optimize objects which we care about, and hence introducing $\mathcal{L}_{adv}^{\mathbf{P}}$ is important to improve visual quality.

Fig. 12: **Ablation Study on Object Discriminator, Object Generator, and Reconstruction Loss.** The first row illustrates panoptic priors utilized as the model's condition. Subsequent rows present synthesized images generated from the same panoptic prior but with various method modifications. These include: omitting the object discriminator (w/o $\mathcal{L}_{adv}^{\mathcal{P}}$), treating all objects as stuff (w/o G_θ^{obj}), and incorporating the reconstruction loss (w/ $\mathcal{L}_{recon}$), as initially applied in UrbanGIRAFFE [71].

This can also be seen from the qualitative results in Fig. 12. As shown in the second row of the image, when removing the object discriminator, the cars are of lower quality and sometimes even "disappear" in the rendered images. Due to the absence of object-level supervision, the object generator does not receive clear punishment when predicting zero density throughout the bounding box. This further demonstrates the importance of the object discriminator in learning disentangled scene representations.



Panoptic Prior | w/o $\mathcal{L}_{geo}$ | w/o $\mathcal{L}_{seg}$ | Full

Fig. 13: **Ablation on Semantic and Geometry loss.** The leftmost column presents the panoptic prior as the condition, while the three columns to the right display the rendered RGB, semantic, and depth images from the model without geometry loss (w/o $\mathcal{L}_{geo}$), without the semantic-aware discriminator (w/o $\mathcal{L}_{seg}$), and the full model, respectively. It is apparent that without training with the semantic loss, there is a misalignment between the rendered image and the corresponding semantic, such as vegetation being rendered with the appearance of a building. In contrast, the geometry loss assists the model in learning more fine-grained geometry of various elements.

Object Generator: Next, we remove the object generator G_θ^{obj} and use the stuff generator to represent the full scene except for the background area (w/o G_θ^{obj}), similar to the CC3D [2] approach. This can also be considered as a generative version of GANCraft [22]. In this scenario, there is no longer a G_θ^{obj}, i.e., the instance-level generator of UrbanGen. This means the model can no longer access a shared canonical representation for objects. Meanwhile, the instance bounding boxes are discretized and added to the semantic voxel grid such that objects are also generated by the stuff generator. This implies that the panoptic prior degrades into a semantic prior. As shown in Table 2 and the third-row of Fig. 12, the quality of objects drops significantly. This verifies the importance of both the instance-level prior and the object generator of UrbanGen in decomposing stuff and objects.

Reconstruction Loss: Additionally, we evaluate the impact of the reconstruction loss, originally proposed in our conference version UrbanGIRAFFE but removed in UrbanGen. In UrbanGIRAFFE [71], we observed that using the adversarial loss alone struggles to maintain the semantic meaning of the panoptic prior. Hence, UrbanGIRAFFE uses a reconstruction loss for stuff regions, which is a combination of the MSE loss and perceptual loss l_{vgg} [26]:

$$\mathcal{L}_{recon} = \mathbb{E}\big[\,\big\|\mathbf{M}\odot(\mathbf{I}-\hat{\mathbf{I}})\big\|_2^2 + \lambda_{vgg} l_{vgg}(\mathbf{M}\odot\mathbf{I}, \mathbf{M}\odot\hat{\mathbf{I}})\big]$$

where $\mathbf{I}$ and $\hat{\mathbf{I}}$ are paired samples, and $\mathbf{M}$ denotes a mask that filters out object regions and preserve only stuff regions.

While the reconstruction loss improves the semantic alignment, it degrades the rendering quality, see Table 2 (w/ $\mathcal{L}_{recon}$). This is unsurprising as the reconstruction loss forces the generative model to predict a deterministic "ground truth" image, leading to inferior image fidelity and limiting the diversity of rendered results. To address this issue, UrbanGen introduces a semantic-aware discriminator to achieve semantic alignment between the rendered image and the conditioned panoptic prior, without satisfying the rendering quality. Note that UrbanGen w/ reconstruction loss still slightly outperforms our conference version thanks to the improved backbone architecture, including larger-capacity upsampler and anti-aliasing design following EG3D [10].

As shown in the fourth-row of Fig. 12, compared to UrbanGen, the model with reconstruction loss tends to produce more artifacts and less detail. UrbanGen achieves better image fidelity while maintaining good semantic alignment using the semantic alignment loss as discussed in the following.

Semantic Alignment Loss: We now evaluate the effectiveness of the semantic-aware discriminator. Table 2 (w/o $\mathcal{L}_{seg}$) shows that the semantic-aware discriminator and the corresponding semantic alignment loss have minimal impact on the rendering fidelity and geometry quality. To better assess the semantic alignment performance, we compare the semantic map obtained from volume rendering to a semantic map predicted by a pre-trained segmentation model [59]. We then report the pixel-level accuracy between these two maps for several major semantic categories, including road, sidewalk, terrain, vegetation, building, and wall, as well as the average pixel accuracy (Acc) of the full images. As reported in Table 3, our approach significantly outperforms both the method without semantic loss (w/o $\mathcal{L}_{seg}$) and the alternative using a reconstruction loss (w $\mathcal{L}_{recon}$), which was adopted in UrbanGIRAFFE.

Furthermore, as shown in Fig. 13, the semantic-aware discriminator helps the model better align the rendered RGB image and the corresponding semantic. As indicated by the red arrows, when training without semantic loss, the model incorrectly generates an appearance similar to a building in areas corresponding to trees. In contrast, the full model effectively achieves semantic-appearance alignment.

Geometry loss: Finally, we exclude the geometry loss $\mathcal{L}_{geo}$ as shown in Table 2 (w/o $\mathcal{L}_{geo}$). This leads to larger depth error and worses the overall geometry quality of the scene. Additionally, the geometry loss contributes to better object fidelity, i.e., lower $\text{FID}_\mathbf{P}$, indicating that the geometry loss helps in learning better object generation. As illustrated in Fig. 13, the geometry loss aids the model in learning more

	Road	Sdwlk	Terr	Vegt	Bldg	Wall	Acc
w $\mathcal{L}_{recon}$	**92.7**	66.2	77.1	83.0	87.6	72.1	74.0
w/o $\mathcal{L}_{seg}$	89.8	72.3	69.3	69.0	72.2	64.5	70.2
Full	91.1	**85.3**	**81.4**	**86.5**	**90.1**	**89.6**	**84.8**

TABLE 3: **Ablation Study on Semantic Alignment.** We report semantic alignment between render semantics and synthesis image on KITTI-360 with different method variations, including adding reconstruction loss (w $\mathcal{L}_{recon}$), without semantic-aware discriminator (w/o $\mathcal{L}_{seg}$).

Method	KITTI-360		Waymo		NuScenes	
	$\text{FID}_\mathbf{I}\downarrow$	$\text{FID}_\mathbf{P}\downarrow$	$\text{FID}_\mathbf{I}\downarrow$	$\text{FID}_\mathbf{P}\downarrow$	$\text{FID}_\mathbf{I}\downarrow$	$\text{FID}_\mathbf{P}\downarrow$
· Specific only	**6.1**	32.2	14.5	55.1	37.6	88.7
Unified only	8.2	29.0	11.2	29.9	25.5	37.0
Unified + Specific	6.9	**24.8**	**8.2**	**29.0**	**21.4**	**35.6**

TABLE 4: **Ablation on Training Strategy.** Specific Only: Training separate models, each trained on a specific dataset. Unified Only: Training solely on a unified dataset. Unified + Specific: First pretraining on a unified dataset followed by fine-tuning on a specific dataset. The evaluation metrics $\text{FID}_\mathbf{I}$ and $\text{FID}_\mathbf{P}$ ($\times 10^3$) are reported to assess the fidelity of full images and object patches.

fine-grained geometry of certain detailed elements, such as the shape of poles. Moreover, the geometry loss facilitates the background generation, where no geometry information is provided by the panoptic prior. Incorporating geometry loss leads to fewer artifacts and more reasonable geometry in the background areas.

Training Strategy: Lastly, we conduct an ablation study on the training strategy, comparing the following methods: 1) Specific Only: Training seperate models, each exclusively trained on a specific dataset, 2) Unified Only: Training solely on the unified urban dataset, and 3) Unified + Specific: Initially training on the unified urban dataset followed by fine-tuning on a specific dataset. As depicted in Table 4, our proposed "Unified + Specific" strategy achieves the best image fidelity. Due to the low resolution and significant occlusions of vehicles in some datasets, models trained in the "Specific Only" setting struggles to learn effective object generation, particularly on Waymo and nuScenes. By training in the "Unified Only" setting, the object fidelity is significantly improved. Notably, on the relatively small-scale nuScenes dataset, training on the unified dataset significantly enhances the rendering quality for both the full image and the object patches. Furthermore, by combining this with specific dataset fine-tuning, our "Unified + Specific" strategy further enhances the fidelity of each dataset.

5 Conclusion

We propose UrbanGen to tackle controllable 3D-aware image synthesis for challenging urban scenes. By effectively incorporating 3D panoptic prior, our model decomposes the scene into stuff, objects, and sky. Our compositional generative model enables diverse controllability regarding large camera viewpoint change, semantic layout, object manipulation, and scene style editing. We believe that our method pushes the frontier of 3D-aware generative models for unbounded scenes with complex geometry.

Acknowledgements

This work is supported by NSFC under grant 62202418 and U21B2004. Yiyi Liao is with the Zhejiang Provincial Key Laboratory of Information Processing, Communication and Networking (IPCAN), Hangzhou, China. Andreas Geiger was supported by the ERC Starting Grant LEGO-3D (850533) and the DFG EXC number 2064/1 - project number 390727645.

Yuanbo Yang is a master's student at Zhejiang University. Before that, he received his B.S. degree at Hangzhou Dianzi University. His research interests include 3D computer vision and generative models.

Yujun Shen is a senior staff research scientist at Ant Research. Before that, he worked as a senior researcher at ByteDance Inc. He received his Ph.D. degree at the Chinese University of Hong Kong and his B.S. degree at Tsinghua University. His research interests include computer vision and deep learning, particularly in 3D vision and generative models. He is an award recipient of Hong Kong PhD Fellowship.

Yue Wang received the Ph.D. degree from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2016. He is currently working as a Professor with the Department of Control Science and Engineering, Zhejiang University. His current research interests include autonomous robots and robot learning.

Andreas Geiger received his Diploma in computer science and his Ph.D. degree from Karlsruhe Institute of Technology in 2008 and 2013. Currently, he is leading the Autonomous Vision Group at the University of Tubingen. He is also a core faculty member of the Tübingen AI Center. His research interests include computer vision, machine learning and scene understanding with a focus on self-driving vehicles.

Yiyi Liao is an assistant professor at Zhejiang University, leading the X-Dimensional Representations Lab. She received her Ph.D. in Control Science and Engineering from Zhejiang University in June 2018 and her B.S. degree from Xi'an Jiaotong University in 2013. Her research interests include 3D vision and scene understanding.

REFERENCES

[1] Waymo open dataset: An autonomous driving dataset. https://www.waymo.com/open, 2019. 4, 9

[2] Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Xingguang Yan, Gordon Wetzstein, Leonidas J. Guibas, and Andrea Tagliasacchi. CC3D: layout-conditioned generation of compositional 3d scenes. *arXiv.org*, 2023. 3, 7, 10, 11, 12, 14

[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 Josh M. Susskind. GAUDI: A neural architect for immersive 3d scene generation. *arXiv.org*, 2022. 3

[4] Alexander W. Bergman, Petr Kellnhofer, Yifan Wang, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. *arXiv.org*, 2022. 3

[5] Mikolaj Binkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2018. 9

[6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv.org*, abs/2311.15127, 2023. 3

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

[8] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 4, 9

[9] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2nerf: Unsupervised conditional π-gan for single image to neural radiance fields translation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[10] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 8, 10, 11, 12, 14

[11] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[12] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *arXiv.org*, 2023. 1

[13] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Scenedreamer: Unbounded 3d scene generation from 2d image collections. *arXiv.org*, 2023. 3, 10

[14] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 9

[15] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. GRAM: generative radiance manifolds for 3d-aware image generation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[16] Terrance DeVries, Miguel Ángel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 3

[17] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proc. Conf. on Robot Learning (CoRL)*, 2017. 1

[18] Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. MagicDrive3D: Controllable 3d generation for any-view rendering in street scenes. *arXiv.org*, 2024. 3

[19] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. MagicDrive: Street view generation with diverse 3d geometry control. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2024. 3

[20] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv.org*, 2024. 3

[21] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2022. 3

[22] Zekun Hao, Arun Mallya, Serge J. Belongie, and Ming-Yu Liu. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 3, 7, 14

[23] Philipp Henzler, Niloy J Mitra, , and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 3

[24] 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 *Advances in Neural Information Processing Systems (NIPS)*, 2017. 9

[25] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. EVA3D: compositional 3d human generation from 2d image collections. *arXiv.org*, 2022. 3

[26] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2016. 14

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 4

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 10, 11

[29] Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, and Sanja Fidler. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[30] Jumin Lee, Sebin Lee, Changho Jo, Woobin Im, Juhyeong Seon, and Sung-Eui Yoon. Semcity: Semantic scene generation with triplane diffusion. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[31] Minsoo Lee, Chaeyeon Chung, Hojun Cho, Min-Jung Kim, Sanghun Jung, Jaegul Choo, and Minhyuk Sung. 3d-gif: 3d-controllable object generation via implicit factorized representations. *arXiv.org*, 2022. 3

[32] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, Yue Wang, Hang Zhao, Zhiding Yu, and Chen Feng. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. *arXiv.org*, 2023. 9

[33] Yiyi Liao, Katja Schwarz, Lars M. Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 5

[34] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. In *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2022. 2, 4, 9

[35] Chieh Hubert Lin, Hsin-Ying Lee, Willi Menapace, Menglei Chai, Aliaksandr Siarohin, Ming-Hsuan Yang, and Sergey Tulyakov. Infinicity: Infinite-scale city synthesis. *arXiv.org*, 2023. 3

[36] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems (NIPS)*, 2020. 7

[37] Yuheng Liu, Xinke Li, Xueting Li, Lu Qi, Chongshou Li, and Ming-Hsuan Yang. Pyramid diffusion for fine 3d large scene generation. *arXiv.org*, 2024. 3

[38] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which

training methods for gans do actually converge? In *Proc. of the International Conf. on Machine learning (ICML)*, 2018. 6

[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 *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 7

[40] Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Autorf: Learning 3d object radiance fields from single view observations. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[41] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 3

[42] Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[43] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. *arXiv.org*, 2011.12100, 2020. 1

[44] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *arXiv.org*, volume 2011.12100, 2020. 11

[45] Michael Niemeyer and Andreas Geiger. GIRAFFE: representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 4, 5, 6

[46] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 3

[47] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[48] Ming Qian, Jincheng Xiong, Gui-Song Xia, and Nan Xue. Sat2density: Faithful density learning from satellite-ground image pairs. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 3

[49] Kyle Sargent, Jing Yu Koh, Han Zhang, Huiwen Chang, Charles Herrmann, Pratul Srinivasan, Jiajun Wu, and Deqing Sun. VQ3D: Learning a 3D-aware generative model on ImageNet. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 3

[50] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021. 5

[51] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021. 7

[52] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems (NIPS)*, 2020. 3

[53] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[54] Zifan Shi, Sida Peng, Yinghao Xu, Yiyi Liao, and Yujun Shen. Deep generative models on 3d representations: A survey. *arXiv.org*, 2022. 1

[55] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. IDE-3D: interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Trans. on Graphics*, 2022. 3

[56] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[57] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[58] Feitong Tan, Sean Fanello, Abhimitra Meka, Sergio Orts-Escolano, Danhang Tang, Rohit Pandey, Jonathan Taylor, Ping Tan, and Yinda Zhang. Volux-gan: A generative model for 3d face synthesis with HDRI relighting. In *ACM Trans. on Graphics*, 2022. 3

[59] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv.org*, 2020. 9, 14

[60] Ayush Tewari, MalliKarjun B R, Xingang Pan, Ohad Fried, Maneesh Agrawala, and Christian Theobalt. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[61] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. In *Advances in Neural Information Processing Systems (NIPS)*, 2023. 9

[62] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 10, 11, 12

[63] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv.org*, 2023. 3

[64] Zhennan Wu, Yang Li, Han Yan, Taizhang Shang, Weixuan Sun, Senbo Wang, Ruikai Cui, Weizhe Liu, Hiroyuki Sato, Hongdong Li, and Pan Ji. Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. *ACM Trans. on Graphics*, 43(4), 2024. 3

[65] Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, and Ziwei Liu. CityDreamer: Compositional generative model of unbounded 3D cities. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[66] Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, and Ziwei Liu. GaussianCity: Generative gaussian splatting for unbounded 3D city generation. *arXiv.org*, 2024. 3

[67] Yinghao Xu, Menglei Chai, Zifan Shi, Sida Peng, Ivan Skorokhodov, Aliaksandr Siarohin, Ceyuan Yang, Yujun Shen, Hsin-Ying Lee, Bolei Zhou, and Sergey Tulyakov. Discoscene: Spatially disentangled generative radiance fields for controllable 3d-aware scene synthesis. *arXiv.org*, 2022. 1, 2, 3, 4, 5, 10, 11, 12

[68] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 11

[69] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. GIRAFFE HD: A high-resolution 3d-aware generative model. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[70] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized predictive model for autonomous driving. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[71] Yuanbo Yang, Yifei Yang, Hanlei Guo, Rong Xiong, Yue Wang, and Yiyi Liao. Urbangiraffe: Representing urban scenes as compositional generative neural feature fields. *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 2, 5, 11, 13, 14

[72] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from A single image. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 7

[73] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3d scene shape from a single monocular image. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2022. 10

[74] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: A 3d generative model for animatable human avatars. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 3

[75] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv.org*, 2010.07492, 2020. 5

[76] Qihang Zhang, Yinghao Xu, Yujun Shen, Bo Dai, Bolei Zhou, and Ceyuan Yang. BerfScene: Generative novel view synthesis with 3D-aware diffusion models. In *CVPR*, 2024. 3

[77] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv.org*, 2024. 3