UrbanGIRAFFE: Representing Urban Scenes as Compositional Generative Neural Feature Fields

Yuanbo Yang, Yifei Yang, Hanlei Guo, Rong Xiong, Yue Wang, Yiyi Liao*

Zhejiang University

{yybbb, yfyang2018, guohl77701, rxiong, ywang24, yiyi.liao}@zju.edu.cn

Abstract

Generating photorealistic images with controllable camera pose and scene contents is essential for many applications including AR/VR and simulation. Despite the fact that rapid progress has been made in 3D-aware generative models, most existing methods focus on object-centric images and are not applicable to generating urban scenes for free camera viewpoint control and scene editing. To address this challenging task, we propose UrbanGIRAFFE, which uses a coarse 3D panoptic prior, including the layout distribution of uncountable stuff and countable objects, to guide a 3D-aware generative model. Our model is compositional and controllable as it breaks down the scene into stuff, objects, and sky. Using stuff prior in the form of semantic voxel grids, we build a conditioned stuff generator that effectively incorporates the coarse semantic and geometry information. The object layout prior further allows us to learn an object generator from cluttered scenes. With proper loss functions, our approach facilitates photorealistic 3D-aware image synthesis with diverse controllability, including large camera movement, stuff editing, and object manipulation. We validate the effectiveness of our model on both synthetic and real-world datasets, including the challenging KITTI-360 dataset.

1. Introduction

Generating photorealistic urban scenes has many applications in simulation, gaming and virtual reality. Unfortunately, designing diverse urban scenes with novel 3D visual content is typically expensive and time-consuming as it requires the expertise of professional artists.

Recent advances in generative models have demonstrated a promising direction to reduce the cost via learning to generate images from data. Ideally, the generated scenes should be controllable in terms of camera pose and 3D con-

Figure 1: **Illustration.** UrbanGIRAFFE generates a photorealistic image given a sampled panoptic prior in the form of a semantic voxel grid and object layout. Our method enables diverse controllability regarding camera pose, instance, and stuff.

tent. For example, the camera should be able to move freely in the scene with six degrees of freedom. The poses of instantiated objects (e.g., cars) should be able to be manipulated independently. Furthermore, the layout of the scene should be controllable.

There are many attempts to generate photorealistic urban images. Several methods study semantic image synthesis to transfer a 2D semantic segmentation map to an RGB urban scene image [25, 53, 57]. However, when changing the camera poses, the generated images across multiple frames may not be consistent using such 2D generative models. Recently, 3D-aware generative models have witnessed rapid progress by lifting the generation process to the 3D space. Despite achieving multi-view consistency, most existing 3D-aware generative models are limited to object-centric images, e.g., faces and cars [58, 10, 9]. There are a few attempts to generate scene images in a compositonal manner [36, 49, 47, 69]. However, all these methods struggle to learn a good geometry of the background and hence

*Corresponding author.

do not support large camera movement, e.g., moving the camera along the road. Another line of work enables camera movement but ignores the compositional nature of the scene, thus lacking controllability of the 3D content [14, 2].

In this paper, we propose UrbanGIRAFFE to address the challenging task of compositional and controllable 3D-aware image synthesis of urban scenes, see Fig. 1. Our key idea is to leverage scene-level but coarse 3D panoptic prior, simplifying the task of learning complex geometry through 2D supervision and incorporating semantic information for scene editing. The panoptic prior, including semantic voxel grids of uncountable stuff and bounding boxes of countable objects, can be obtained from existing datasets [37] or inferred from pre-trained models [8]. Specifically, our model represents the scene as compositional neural feature fields consisting of stuff, objects, and sky. We propose a semantic voxel-conditioned stuff generator, effectively preserving the semantic and geometry information provided by the prior. In terms of objects, we follow GIRAFFE [49] to generate objects in canonical space by leveraging the object layout prior. We further model the sky and far regions using a sky 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. For the complicated urban scenes, we observe that training with an adversarial loss on the full image alone is insufficient. We additionally employ an adversarial loss applied to objects and a reconstruction loss to the stuff image regions to improve the image fidelity.

Our contributions are as follows. i) We propose to study a novel yet challenging task of 3d-aware urban generative models with diverse controllability in terms of large camera movement, object manipulation and stuff editing. ii) We leverage coarse 3D panoptic prior to address this challenging task and design compositional generative radiance fields that leverage the prior information effectively. iii) With our carefully design the training objectives, our method demonstrates state-of-the-art performance compared to existing methods on both synthetic and real-world datasets, including the challenging KITTI-360 dataset.

2. Related Work

Conditional Image synthesis: In recent years, Generative Adversarial Networks [18, 30, 31, 28, 29, 56] have achieved impressive results in photorealistic image synthesis. As it is not straightforward to control the generated images of unconditional GANs, many attempts have been made for conditional image synthesis. A line of works generates images conditioned on a 2D semantic segmentation map [25, 53, 57]. Instead of requiring per-pixel semantic annotation, another line of methods generates images following an image layout in the form of 2D bounding boxes [75, 72, 21] or learned blobs [15]. Recent perpetual view generation methods including Infinite Nature [39], InfiniteNature-Zero [35] and DiffDreamer [6] are trained on nature videos or photo collections and demonstrate the capability to generate unbounded videos of natural scenes. However, when changing the camera viewpoint, the generated images across different views are typically not multi-view consistent, as discussed in [20]. We instead learn a 3D-aware conditional generative model that leads to better consistency with the underlying 3D representation.

3D-Aware Image Synthesis: 3D-aware generative models have received growing attention recently. While early works learn to generate 3D voxel grids [46, 22], recent methods achieve high-fidelity 3D-aware image synthesis leveraging neural radiance fields as the underlying 3D representation [58, 9, 10, 59, 13, 70, 19]. Empowered by 3D-aware generative models, many promising applications have been demonstrated, including semantic editing [61, 60], relighting [62, 34], single-view reconstruction [7, 44] and articulated human generation [74, 50, 4, 24]. However, all aforementioned methods focus on object-centric scenes and assume that the object lies in a canonical object coordinate system. Thus, it is non-trivial to extend these methods to complex, unaligned urban scenes. GSN [14] and GAUDI [3] propose to generate unbounded indoor scenes. However, both ignore the compositionality of the scene, thus making it harder to achieve high visual fidelity and do not support editing of the scene content.

A few works exploit the compositionality of 3D scenes to generate scenes containing multiple objects [36, 69, 47, 49, 71]. These methods, including GIRAFFE [49], consider the compositionality of foreground objects only, thus are incapable of modeling complex background geometry in urban scenes. A concurrent work, DiscoScene [69], also studies 3D-aware generative model of urban scenes. Despite achieving high-fidelity image synthesis, DiscoScene does not support camera control or stuff editing in urban scenes. Several other concurrent works explore diverse aspects of scene-level 3D generation. CC3D [1] focuses on synthesizing complex 3D scenes based on 2D semantic scene layouts of foreground objects. SceneDreamer [12] generates high-fidelity infinite natural landscapes, yet it is non-trivial to be extended to complex urban scenes. InfiniCity [38] achieves infinite city generation by lifting synthesized 2D layout to 3D based on test-time optimization, thus requiring a long optimization time for generating a new scene.

Neural Radiance Fields: We proposed to present the scene as compositional neural feature fields. Exploiting implicit neural representations [42, 52], NeRF [43] has enabled impressive novel view synthesis by training a single model for each scene. Many exciting works have shown its potential in real-time rendering [54, 45], geometric recon-
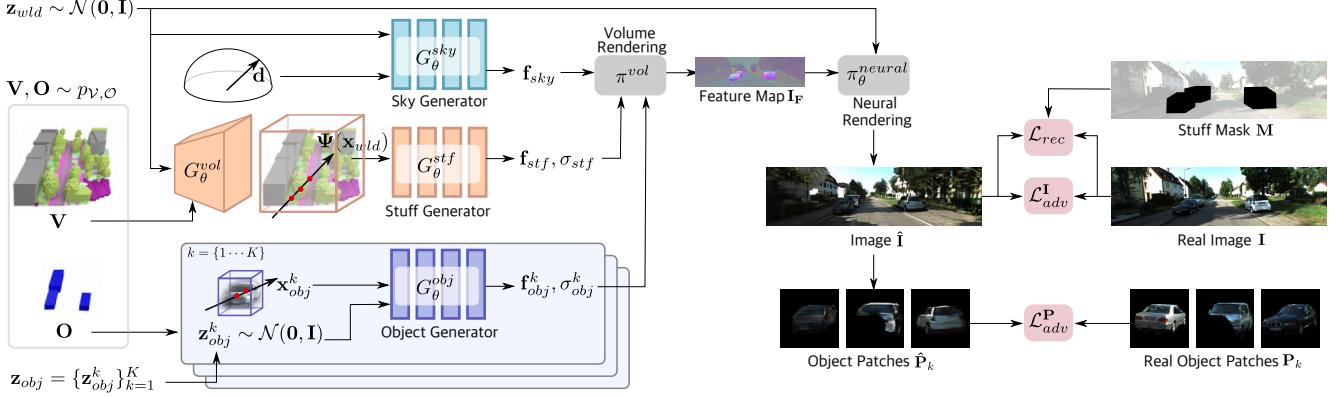
Figure 2: **Method Overview.** We leverage panoptic prior in the form of semantic voxel grids and instance object layouts to build a 3D-aware generative model for urban scenes. Our model takes as input a global noise vector $\mathbf{z}_{wld}$ for the entire scene, K noise vectors $\{\mathbf{z}_{obj}^k\}_{k=1}^K$ for objects, and a sampled panoptic prior $\mathbf{V}, \mathbf{O} \sim p_{\mathcal{V}, \mathcal{O}}$. We decompose the scene into sky, stuff, and objects. The stuff generator is conditioned on the semantic voxel grid $\mathbf{V}$ to preserve its semantic and geometry information. The objects are generated in the canonical object coordinate system guided by $\mathbf{O}$. Combined with the sky generator, a feature map $\mathbf{I_F}$ is obtained via volume rendering. We further leverage neural rendering to output the RGB image $\hat{\mathbf{I}}$ and object patches $\hat{\mathbf{P}}_k$. The full 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 reconstruction loss $\mathcal{L}_{rec}$ for stuff regions.

struction [64, 66], semantic segmentation [76, 16, 33], and view synthesis from sparse input [65, 73, 11]. It has been shown that NeRF can also be extended to model unbounded urban scenes [55, 51] and scale to city level [63, 67, 68].

While all these methods focus on reconstructing existing scenes, we aim to learn a conditional generative model that can generate urban images conditioned on different panoptic layouts. A more related work, GANCraft [20], aims to generate a scene based on semantic voxels, yet it also requires per-scene optimization. In contrast, our generative model allows for stuff editing by manipulating the semantic voxels.

3. Method

In UrbanGIRAFFE, our goal is to build compositional generative feature 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 uncountable stuff, countable objects, and sky, see Fig. 2 for an overview. We assume prior distributions are provided for both stuff and objects in order to disentangle the complicated urban scenes. Given a camera pose, we render a composited feature map and generate the target image via neural rendering. Our model is trained end-to-end with adversarial and reconstruction losses.

In this section, we first introduce the prior distributions of stuff and objects, respectively. Next, we introduce our compositional generative model for urban scene generation. Finally, we describe the sampling strategy, loss functions and implementation details.

3.1. Panoptic Prior

We assume a prior distribution of the scene layout is given in order to train our 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 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 dataset [37] to form the distribution $p_{\mathcal{O}}$. This distribution can 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 follows the idea of GIRAFFE [49] which represents the urban scene as a compositional neural feature fields. A key difference is that we model the background using a stuff generator conditioned on a semantic voxel grid, assisted with a sky generator to model sky and far regions. The stuff and the sky generator share a global latent code $\mathbf{z}_{wld} \in \mathcal{N}(\mathbf{0}, \mathbf{I})$, whereas each object has its own latent code $\mathbf{z}_{obj} = \{\mathbf{z}_{obj}^k \in \mathcal{N}(\mathbf{0}, \mathbf{I})\}_{k=1}^K$ to ensure the diversity of object shape and appearance in a scene. We now describe each of these generators in detail.

Object Generator: For objects, we follow existing compositional methods to generate each object k in a normalized object coordinate space [36, 49]:

$$G_\theta^{obj} : (\gamma(\mathbf{x}_{obj}^k), \mathbf{z}_{obj}^k) \mapsto (\mathbf{f}_{obj}^k, \sigma_{obj}^k) \qquad (1)$$

where G_θ^{obj} denotes the object generator that maps a 3D point $\mathbf{x}_{obj}^k$ encoded by positional encoding $\gamma(\cdot)$ and a noise vector $\mathbf{z}_{obj}^k$ to a feature vector $\mathbf{f}_{obj}^k \in \mathbb{R}^{M_f}$ and density σ_{obj}^k. 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 $\{\mathbf{R}, \mathbf{t}, \mathbf{s}\}$.

$$\mathbf{x}_{wld} = \mathbf{R}(\mathbf{s} \odot \mathbf{x}_{obj}^k) + \mathbf{t} \qquad (2)$$

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.

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 [53, 57], 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 noise vector $\mathbf{z}_{wld}$ to a feature grid $\boldsymbol{\Psi} \in \mathbb{R}^{H_v \times W_v \times D_v \times M_v}$ conditioned on the semantic voxel grid $\mathbf{V} \in \mathbb{R}^{H_v \times W_v \times D_v \times L}$:

$$G_\theta^{vol} : (\mathbf{z}_{wld}, \mathbf{V}) \mapsto \boldsymbol{\Psi} \qquad (3)$$

In practice, G_θ^{vol} is a 3D convolutional neural network. The semantic condition $\mathbf{V}$ is injected at multiple resolutions using spatially-adaptive normalization, see Fig. 3 as an illustration. Given a 3D point $\mathbf{x}_{wld}$, we trilinearly interpolate a feature vector $\boldsymbol{\Psi}(\mathbf{x}_{wld}) \in \mathbb{R}^{M_v}$. Next, we map $\mathbf{x}_{wld}$ and $\boldsymbol{\Psi}(\mathbf{x}_{wld})$ to the final stuff feature $\mathbf{f}_{stf} \in \mathbb{R}^{M_f}$ and density σ_{stf} using the MLP head:

$$G_\theta^{stf} : (\boldsymbol{\Psi}(\mathbf{x}_{wld}), \gamma(\mathbf{x}_{wld})) \mapsto (\mathbf{f}_{stf}, \sigma_{stf}) \qquad (4)$$
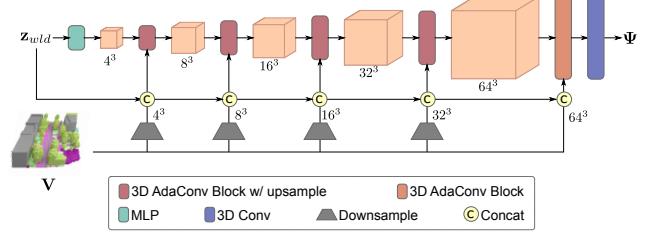

Figure 3: **Feature Grid Generator** G_θ^{vol} as a part of the stuff generator. We adopt spatially-adaptive normalization to inject the semantic condition $\mathbf{V}$ and the noise vector $\mathbf{z}_{wld}$ at multiple resolutions.

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.

Sky Generator: The stuff generator cannot model regions far from the semantic voxel grid, e.g., sky. Therefore, we model the sky and other far regions as an infinitely far away dome following [20, 55]. Specifically, we use a sky generator G_θ^{sky} to map a ray direction $\mathbf{d}$ to a sky feature vector $\mathbf{f}_{sky} \in \mathbb{R}^{M_f}$.

$$G_\theta^{sky} : (\mathbf{z}_{wld}, \mathbf{d}) \mapsto \mathbf{f}_{sky}$$

Note that the global latent code $\mathbf{z}_{wld}$ is used to ensure the style consistency between sky and other semantics within an urban scene.

Compositional Volume Rendering: We accumulate feature vectors of objects, stuff, and sky on each ray via compositional volume rendering. We first sample points from the object and stuff generators independently (the sampling strategy will be elaborated in Section 3.3). Next, we sort all points wrt. their distances to the camera center and accumulate their feature vectors via volume rendering. Finally, the sky feature is added to non-opaque regions.

Formally, let $\{\mathbf{x}_i\}_{i=1}^M$ denote M sorted points on a ray, compositing of $\mathbf{x}_{wld}$ sampled for the stuff generator and $\mathbf{x}_{obj}^k$ sampled for the object generators (transformed to the world coordinate system via Eq. 2). $\mathbf{f}_i$ and σ_i denote the corresponding feature vector and density at $\mathbf{x}_i$. The volume rendering is

$$\pi^{vol} : \{\mathbf{f}_i, \sigma_i, \mathbf{f}_{sky}\}_{i=1}^M \mapsto \mathbf{F} \qquad (5)$$

Specifically, $\mathbf{F}$ is obtained via numerical integration as

$$\mathbf{F} = \sum_{i=1}^N T_i \alpha_i \mathbf{f}_i + (1 - \sum_{i=1}^N T_i \alpha_i) \mathbf{f}_{sky} \qquad (6)$$

$$\alpha_i = 1 - e^{(-\sigma_i \delta_i)} \quad T_i = \prod_{j=1}^{i-1}(1 - \alpha_j) \qquad (7)$$

where T_i and α_i denote transmittance and alpha value of a sample point $\mathbf{x}_i$.

2D Neural Rendering: Following [49], 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 $\mathbf{I_F} \in \mathbb{R}^{H_f \times W_f \times M_f}$ and the noise vector $\mathbf{z}_{wld}$ to the RGB image $\hat{\mathbf{I}} \in \mathbb{R}^{H \times W \times 3}$ at the target resolution. Here, $\mathbf{z}_{wld}$ is adopted to enable content-aware upsampling.

$$\pi_\theta^{neural} : (\mathbf{I_F}, \mathbf{z}_{wld}) \mapsto \hat{\mathbf{I}} \tag{8}$$

Object Patch Rendering: In addition to the full image $\hat{\mathbf{I}}$, we further generate a set of object patches, see Fig. 2 as an illustration. We upsample object masks obtained from volume rendering to segment the objects after neural rendering. Please refer to the supplementary for more details.

3.3. 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.

Ray-Voxel Intersection Sampling for Stuff: Inspired by existing methods [20, 40], 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 hit, and then sample M_{vol} points within 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 [43], i.e., a random shift is added to the sampled points.

3.4. Loss Functions

We train the entire model end-to-end using adversarial training aided by a reconstruction loss for stuff regions.

Adversarial Loss: We apply an adversarial to the composited image. Let G_θ denote the full conditional generator that maps the noise vectors and the panoptic prior to a full RGB image:

$$G_\theta : (\mathbf{z}_{wld}, \mathbf{z}_{obj}, \mathbf{V}, \mathbf{O}) \mapsto \hat{\mathbf{I}} \tag{9}$$

We apply the non-saturated adversarial loss with R1-regularization [41]:

$$\mathcal{L}_{adv}^{\mathbf{I}} = \mathbb{E}_{\mathbf{I} \sim p_\mathcal{D}} \left[f(-D_\phi^{\mathbf{I}}(\mathbf{I})) - \lambda \| \nabla D_\phi^{\mathbf{I}}(\mathbf{I}) \|^2 \right] +$$
$$\mathbb{E}_{\mathbf{z}_{wld}, \mathbf{z}_{obj} \sim \mathcal{N}, \mathbf{V}, \mathbf{O} \sim p_{\mathcal{V}, \mathcal{O}}} \left[f(D_\phi^{\mathbf{I}}(G_\theta(\mathbf{z}_{wld}, \mathbf{z}_{obj}, \mathbf{V}, \mathbf{O}))) \right] \tag{10}$$

Note that the visual quality of objects like cars is essential for urban scenes. Unfortunately, objects do not always occupy a large area in urban images. Our experiments show that using scene-level adversarial training alone fails to generate photorealistic objects. Inspired by existing methods [17, 69], 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 $\mathcal{L}_{adv}^{\mathbf{P}}$ similar to $\mathcal{L}_{adv}^{\mathbf{I}}$.

Stuff Reconstruction Loss: For our conditional stuff generator, we observe that using adversarial loss alone struggles to generate photorealistic results. One possible reason is that learning generative 3D feature fields for complex stuff regions is more challenging than the object-centric generation. To stabilize adversarial training and improve the quality of synthesized images, we further leverage reconstruction loss for stuff regions. Following [20], our reconstruction loss is a combination of the MSE loss and perceptual loss l_{vgg} [26]:

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

where $\mathbf{I}$ and $\hat{\mathbf{I}}$ are paired samples, and $\mathbf{M}$ denotes a mask that filters out object regions based on the projected 3D projecting boxes $\mathbf{O}$. Since our stuff generator is a conditioned generative model depending on the semantic voxel grid, adding the reconstruction loss is reasonable as the appearance is highly relevant to the corresponding semantic label. This provides stronger supervision that $\mathbf{z}_{wld}$ only needs to model the variation within the same semantic class.

3.5. Implementation Details

We use 3D CNNs with 5 spatially-adaptive normalization blocks for the feature grid generator G_θ^{vol}. We set $H_v = W_v = D_v = 64$ for all experiments, i.e., the semantic voxel grids are at the resolution of 64^3. We use $M_v = 16$ channels for the feature grid $\mathbf{\Psi}$ to avoid large memory consumption. The MLP head G_θ^{stf} of the stuff generator is an 8-layer ReLU MLP with a hidden dimension of 256. The object generator G_θ^{obj} is also an 8-layer ReLU MLP with a hidden dimension of 128. In terms of the sky generator G_θ^{sky}, a 5-layer MLP with a hidden dimension of 256 is adopted. All these MLP generators output feature vectors of dimension $M_f = 32$.

	KITTI-360		CLEVR-W	
	FID$_\mathbf{I}$ ↓	KID$_\mathbf{I}$ ↓	FID$_\mathbf{I}$ ↓	KID$_\mathbf{I}$ ↓
2D GAN [32]	32.2	0.020	17.7	0.019
GSN [14]	127.1	0.124	–	–
GIRAFFE [49]	110.9	0.122	103.9	0.101
Ours	**43.3**	**0.040**	**25.7**	**0.019**

Table 1: **Quantitative Comparison** on KITTI-360 and CLEVR-W. Our method outperforms 3D-aware baseline methods and is comparable to the 2D baseline.

During training, we sample camera poses along plausible driving trajectories given a semantic voxel grid. Regarding ray marching, we sample $M_{obj} = 12$ points within each object's bounding box and $M_{vol} = 6$ within each voxel. We use the Adam optimizer with a batch size of 16. The learning rates of the discriminator and the generator are 1×10^{-4} and 2×10^{-4}, respectively. During inference, we generate images using a moving averaged model with an exponential decay of 0.999 for the weights.

4. Experimental Results

In this section, we first compare our method to several 2D and 3D baselines on both synthetic and real-world datasets. Subsequently, we design a number of controllable urban scene editing experiments to evaluate the preferences of our synthesis model with regards to controllability and fidelity. We further conduct ablation studies to better understand the influence of different architectural components and report runtime analysis.

Datasets: We conduct experiments on two multi-object datasets with diverse backgrounds. **KITTI-360** [37] is an outdoor sub-urban dataset containing complex scene geometry. Furthermore, scenes in KITTI-360 are replete with highlights and shadows, causing the appearance of the same object to vary greatly in different scenes. KITTI-360 provides coarse 3D bounding primitives in cuboids and spheres for both stuff and objects. We consider cars as objects since cars are important for driving scenarios. For stuff regions, we simply convert the coarse 3D bounding primitives to semantic voxel grids. We further create an augmented **CLEVR-W** dataset following CLEVR [27]. In contrast to existing methods [49, 69] that places objects on a simple flat background in CLEVR, we introduce walls into the background. We consider the wall and the floor as stuff regions. Please refer to the supplementary material for additional information regarding the CLEVR-W dataset.

Baselines: We compare our approach to two state-of-the-art models GIRAFFE [49] and GSN [14] for 3D-aware image synthesis. To further evaluate the fidelity of the synthesized image, we additionally compare our method with a

state-of-the-art 2D method, StyleGAN2 [31].

Metrics: We report the FID [23] and KID [5] scores to quantify image quality. We use 5k real and fake samples to calculate the FID and KID scores. We split the panoptic priors into training and validation for evaluation as our training objectives include a reconstruction loss. For the KITTI-360 dataset, we split it into windows of 200 meters without overlapping and sample 20% windows for validation. Specifically, we randomly sample semantic voxel grids for synthesizing fake images and other images of different layouts as real ones. Both semantic voxel grids and real images are randomly sampled from the unseen validation windows. This approach ensures that the evaluation assesses the model's performance on previously unseen layouts, enabling us to gauge its generalization ability. In addition to the window-based split, we further keep an entire sequence from the KITTI-360 dataset to create the validation set. This sequence-based split introduces a more challenging evaluation setting for our model. The details can be found in the supplementary material.

4.1. Comparison to the State of the Art

Quantitative Comparison: Table 1 shows the quantitative comparison on KITTI-360 and CLEVR-W. Note that GSN requires training on sequential frames, thus we omit GSN on the CLEVR-W dataset which does not contain sequential data. The quantitative comparison shows that our method greatly outperforms existing state-of-the-art 3D methods regarding image fidelity and is comparable to the 2D baseline.

Qualitative Comparison: We compare our method with GIRAFFE and GSN on KITTI-360 in Fig. 4 with the camera moving forward. Note that GIRAFFE struggles to learn the complicated background geometry of urban scenes. This distracts the GAN training, thus leading to low-quality results even in a static scenario (the first row). Compared with GIRAFFE, GSN's scene representation is built upon a local 2D feature map, enabling it to model relatively complex 3D scenes. Therefore, GSN performs better in the static scenario, but the image quality drops dramatically as the camera moves forward. As a comparison, our method is conditioned on a 3D semantic voxel grid, thus enabling photorealistic and consistent 3D-aware image synthesis even with a large camera moving distance.

Fig. 5 shows the qualitative comparison with GIRAFFE on CLEVR-W. We conduct various experiments including stuff editing (e.g., editing the height of the wall or moving it closer to the objects), object rearrangement, and camera viewpoint manipulation. Note that GIRAFFE performs well on the foreground objects but still lags behind on the background. In contrast, our method can keep high fidelity and 3D consistency under these experiments, which clearly outperforms the baseline method.

GIRAFFE [49]   GSN [14]   UrbanGIRAFFE (Ours)

Figure 4: **Qualitative Comparison on KITTI-360**. The 1st and 5th rows show images rendered at the default camera pose for each method. The camera moves forward in the following rows, with an accumulated moving distance of 10 meters.
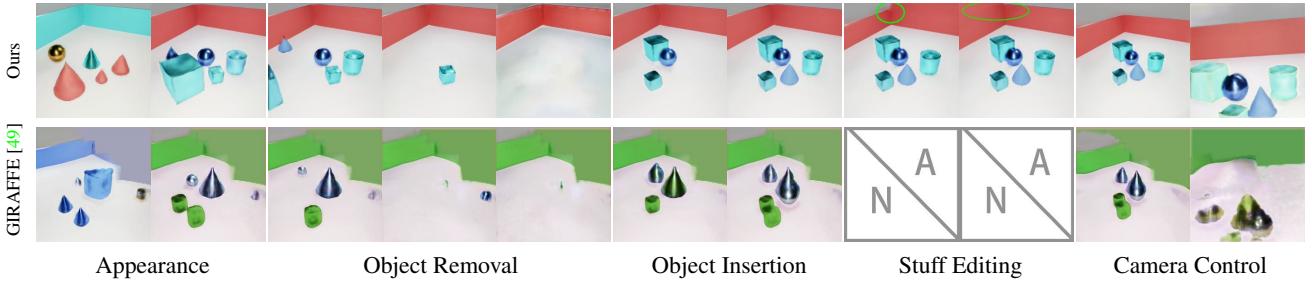

Appearance   Object Removal   Object Insertion   Stuff Editing   Camera Control

Figure 5: **Qualitative Comparison on CLEVR-W**. We compare with GIRAFFE wrt. various controllable image synthesis tasks. Our method outperforms GIRAFFE in modeling the background, thus enabling stuff editing and better performance in camera viewpoint control.

4.2. Controllable Urban Scene Generation

We now demonstrate the diverse controllability of our model in terms of stuff editing, object editing and camera viewpoint control.

Stuff Editing: Our semantic-conditioned stuff generator enables fine-grained stuff editing by modifying the conditioning semantic voxel. As shown in Fig. 6a. We can transfer stuff semantics like "Road to Grass" and "Building to Tree". It is also possible to edit the occupancy of the voxel
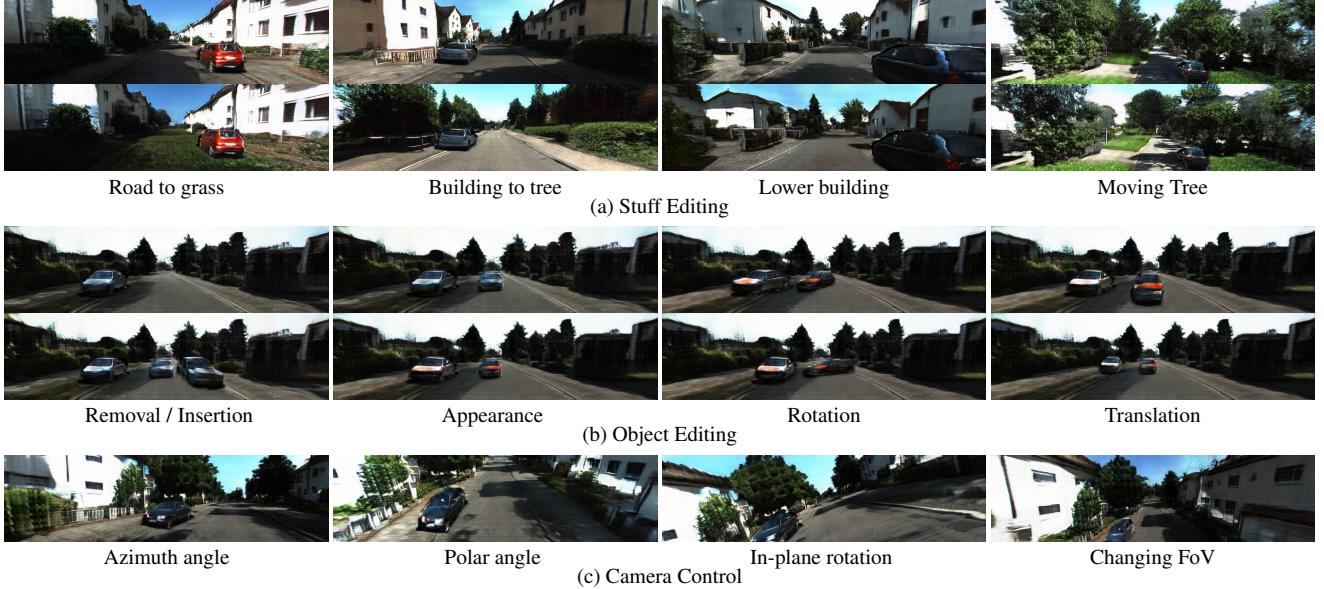
Road to grass   Building to tree   Lower building   Moving Tree

(a) Stuff Editing

Removal / Insertion   Appearance   Rotation   Translation

(b) Object Editing

Azimuth angle   Polar angle   In-plane rotation   Changing FoV

(c) Camera Control

Figure 6: **Controllable 3D-aware Image Synthesis** on KITTI-360.

w/o $\mathcal{L}_{recon}$ w/o $\mathcal{L}_{adv}^{\mathbf{P}}$ w/o G_θ^{obj} Full Model

Figure 7: **Ablation Study**. Each row shows synthesized images given the same panoptic prior in KITTI-360 with different method variations. Removing the reconstruction loss (w/o $\mathcal{L}_{recon}$) leads to more artifacts, and the semantic condition may not be preserved (3rd row). Removing the object discriminator (w/o $\mathcal{L}_{adv}^{\mathbf{P}}$) and modeling all objects as stuff (w/o G_θ^{obj}) both impede the fidelity of objects (cars).

grids, e.g., "Lower building" and "Move tree". All these stuff editings are achieved by modifying the semantic voxel grid without additional optimization.

It is worth mentioning that, in the "Building To Tree" example, the shadow of the road also changes to a large degree after the editing. This suggests that our method not only allows for photorealistic and semantically-align urban scene generation but also learns the implicit relationship between the shadow condition and semantic layout.

Object Editing: Next, we conduct various experiments on object editing in Fig. 6b. As in GIRAFFE [48], we can add/delete objects, and control their appearance, rotation, and translation. Our object experiments with object editing do not affect the appearance of other scene parts, suggesting that our method can disentangle objects from the complex background by leveraging the panoptic prior.

Camera Control: Finally, Fig. 6c shows that 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.

4.3. 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.

Reconstruction Loss: We first validate the role of the reconstruction loss. After removing the reconstruction loss, the FID and KID scores drop significantly (w/o $\mathcal{L}_{recon}$). This is unsurprising as the reconstruction loss provides stronger supervision to align the generated scenes with the ground truth. Fig. 7 shows that removing the reconstruction loss can also lead to reasonable performance, but yields

	$\text{FID}_\mathbf{I} \downarrow$	$\text{KID}_\mathbf{I} \downarrow$	$\text{FID}_\mathbf{P} \downarrow$	$\text{KID}_\mathbf{P} \downarrow$
w/o $\mathcal{L}_{recon}$	89.3	0.067	77.6	0.062
w/o $\mathcal{L}_{adv}^\mathbf{P}$	53.1	0.050	119.0	0.120
w/o G_θ^{obj}	44.7	0.036	–	–
Full	**43.4**	**0.040**	**67.1**	**0.056**

Table 2: **Ablation Study.** We report FID and KID of Urban-GIRAFFE on KITTI-360 without reconstruction loss (w/o $\mathcal{L}_{recon}$), without object discriminator (w/o $\mathcal{L}_{adv}^\mathbf{P}$) and without object generator (w/o G_θ^{obj})

Figure 8: **Diversity**. We compare our synthesized images to the corresponding ground truth image with the same panoptic prior. Note that our method keeps the same layout but maintains diversity.

more artifacts. Moreover, reconstruction loss is particularly important for infrequently encountered semantic classes. For example, removing reconstruction loss results in the model rendering the "rail track" as grass, while the full model can render it with the corresponding semantic meaning faithfully (see 3rd row of Fig. 7). Note that our full model can maintain high fidelity while still exhibiting differences from the ground truth image, see Fig. 8. These findings suggest our full model can produce diverse results instead of simply remembering the entire dataset.

Object Discriminator: Next, 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, removing $\mathcal{L}_{adv}^\mathbf{P}$ significantly increases the patch $\text{FID}_\mathbf{P}$ and $\text{KID}_\mathbf{P}$. This can also be seen from the qualitative results in Fig. 7, where the cars are of lower quality when removing the object discriminator. It is worth noting that $\text{FID}_\mathbf{I}$ is less affected, indicating that in scenes where the proportion of objects pixels is not large, the global adversarial training cannot provide enough supervision to optimize objects which we actually care, and hence introducing $\mathcal{L}_{adv}^\mathbf{P}$ is important to improve visual quality.

All Stuff: Lastly, we remove the object generator G_θ^{obj} and use the stuff generator to represent the full scene except for the sky (w/o G_θ^{obj}), similar to the GSN approach. This can also be considered as a generative version of GAN-Craft [20]. As shown in Fig. 7, the quality of objects drops significantly. This verifies the importance of decomposing stuff and objects to learn high-fidelity object generation. Note that $\mathcal{L}_{adv}^\mathbf{P}$ is also not applied in this experiment as there

	$R_I = 94 \times 352$	$R_I = 188 \times 704$
GSN	184	–
GIRAFFE	133	455
Ours	**96**	**247**

Table 3: **Rendering Time Comparison.** We report time in ms per image on different resolutions. All results are evaluated on a single NVIDIA RTX 3090.

is no information on the object instance.

4.4. Runtime Comparison

Table 3 shows the rendering time comparison of different methods. Our method achieves a rendering speed of approximately 5 frames per second (FPS) for images of resolution 188×704. Our method offers two key advantages in the rendering process. Firstly, in the case of scene rendering, our feature grid is generated only once, allowing subsequent frame renderings to reuse this grid, resulting in improved efficiency. Secondly, our method surpasses the baselines in rendering efficiency by leveraging geometric priors for efficient sampling. This enables us to achieve visually pleasing results while significantly reducing the number of points sampled along each ray. Compared to the baseline methods that require 32 sampled points, our method achieves similar quality with just 4 points sampled for stuff rendering.

5. Conclusion

We propose UrbanGIRAFFE 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, and object manipulation. We believe that our method pushes the frontier of 3D-aware generative models for unbounded scenes with complex geometry. In future work, it can be augmented with a semantic voxel generator for sampling novel scenes. Further, our method does not disentangle light from ambient color, which is worth investigating to enable lighting control.

References

[1] 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. 2023. 2

[2] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[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. 2

[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. 2

[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. 6

[6] Shengqu Cai, Eric Ryan Chan, Songyou Peng, Mohamad Shahbazi, Anton Obukhov, Luc Van Gool, and Gordon Wetzstein. Diffdreamer: Consistent single-view perpetual view generation with conditional diffusion models. *arXiv.org*, 2022. 2

[7] 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. 2

[8] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[9] 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. 1, 2

[10] 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. 1, 2

[11] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 3

[12] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Scenedreamer: Unbounded 3d scene generation from 2d image collections. 2023. 2

[13] 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. 2

[14] Terrance DeVries, Miguel Ángel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2, 6, 7

[15] Dave Epstein, Taesung Park, Richard Zhang, Eli Shechtman, and Alexei A. Efros. Blobgan: Spatially disentangled scene representations. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2

[16] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. 2022. 3

[17] Raghudeep Gadde, Qianli Feng, and Aleix M. Martínez. Detail me more: Improving gan's photo-realism of complex scenes. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 5

[18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 2

[19] 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. 2

[20] 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. 2, 3, 4, 5, 9

[21] Sen He, Wentong Liao, Michael Ying Yang, Yongxin Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Context-aware layout to image generation with enhanced object appearance. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[22] 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. 2

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 6

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

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[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. 5

[27] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[28] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[29] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NIPS*, 2021. 2

[30] 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. 2

[31] 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. 2, 6

[32] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. 2020. 6

[33] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J. Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas A. Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[34] 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. 2

[35] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2

[36] 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. 1, 2, 4

[37] 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, 3, 6

[38] Chieh Hubert Lin, Hsin-Ying Lee, Willi Menapace, Menglei Chai, Aliaksandr Siarohin, Ming-Hsuan Yang, and Sergey Tulyakov. Infinicity: Infinite-scale city synthesis. 2023. 2

[39] Andrew Liu, Ameesh Makadia, Richard Tucker, Noah Snavely, Varun Jampani, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[40] 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. 5

[41] 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. 5

[42] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[43] 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. 2, 5

[44] 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. 2

[45] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv.org*, Jan. 2022. 2

[46] 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. 2

[47] 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. 1, 2

[48] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *arXiv.org*, volume 2011.12100, 2020. 8

[49] 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. 1, 2, 4, 5, 6, 7

[50] 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. 2

[51] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[52] Jeong Joon Park, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation.

In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[53] 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. 1, 2, 4

[54] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[55] Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Thomas A. Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4

[56] Axel Sauer, Katja Schwarz, and Andreas Geiger. Styleganxl: Scaling stylegan to large diverse datasets. In *ACM Trans. on Graphics*, 2022. 2

[57] 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. 1, 2, 4

[58] 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. 1, 2

[59] 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. 2

[60] 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. 2

[61] 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. 2

[62] 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. 2

[63] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben P. Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[64] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NIPS)*, 2021. 3

[65] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 3

[66] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. *arXiv.org*, 2022. 3

[67] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Citynerf: Building nerf at city scale. *arXiv.org*, abs/2112.05504, 2021. 3

[68] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 3

[69] 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, 5, 6

[70] 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. 2

[71] 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. 2

[72] Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. Modeling image composition for complex scene generation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[73] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[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. 2

[75] Bo Zhao, Weidong Yin, Lili Meng, and Leonid Sigal. Layout2image: Image generation from layout. *IJCV*, 2020. 2

[76] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 3