SyntheOcc: Synthesize Geometric-Controlled Street View Images through 3D Semantic MPIs

Leheng Li[1] **Weichao Qiu**[3] **Yingjie Cai**[3] **Xu Yan**[3] **Qing Lian**[2]
Bingbing Liu[3] **Ying-Cong Chen**[1,2*]

HKUST(GZ)[1] HKUST[2] HUAWEI Noah's Ark Lab[3]
Project page: len-li.github.io/syntheocc-web

Abstract

The advancement of autonomous driving is increasingly reliant on high-quality annotated datasets, especially in the task of 3D occupancy prediction, where the occupancy labels require dense 3D annotation with significant human effort. In this paper, we propose **SytheOcc**, which denotes a diffusion model that Synthesize photorealistic and geometric-controlled images by conditioning Occupancy labels in driving scenarios. This yields an unlimited amount of diverse, annotated, and controllable datasets for applications like training perception models and simulation. SyntheOcc addresses the critical challenge of how to efficiently encode 3D geometric information as conditional input to a 2D diffusion model. Our approach innovatively incorporates 3D semantic multi-plane images (MPIs) to provide comprehensive and spatially aligned 3D scene descriptions for conditioning. As a result, SyntheOcc can generate photorealistic multi-view images and videos that faithfully align with the given geometric labels (semantics in 3D voxel space). Extensive qualitative and quantitative evaluations of SyntheOcc on the nuScenes dataset prove its effectiveness in generating controllable occupancy datasets that serve as an effective data augmentation to perception models.

1 Introduction

With the rapid development of generative models, they have shown realistic image synthesis and diverse controllability. This progress has opened up new avenues for dataset generation in autonomous driving [5, 12, 24, 31]. The task of dataset generation is usually modeled as controllable image generation, where the ground truth (*e.g.* 3D Box) is employed to control the generation of new datasets in downstream tasks (*e.g.* 3D detection). This approach helps to mitigate the data collection and annotation effort as it can generate labeled data for free. However, a novel task of vital importance, occupancy prediction [25, 28], poses new challenges for dataset generation compared with 3D detection. It requires finer and more nuanced geometry controllability, which refers to use the occupancy state and semantics of voxels in the whole 3D space to control the image generation. We argue that solving this problem not only allows us to synthesize occupancy datasets, but also empowers valuable applications such as editing geometry to generate rare data for corner case evaluation, as shown in Fig. 1. In the following, we first illustrate why prior work struggles to achieve the above objective, and then demonstrate how we address these challenges.

In the area of diffusion models, several representative works have displayed high-quality image synthesis; however, they are constrained by limited 3D controllability: they are incapable of editing 3D voxels for precise control. For example, BEVGen [24] generates street view images by conditioning BEV layouts using diffusion models. MagicDrive [5] extend BEVGen and additionally converts the

*Corresponding author.

Preprint. Under review.

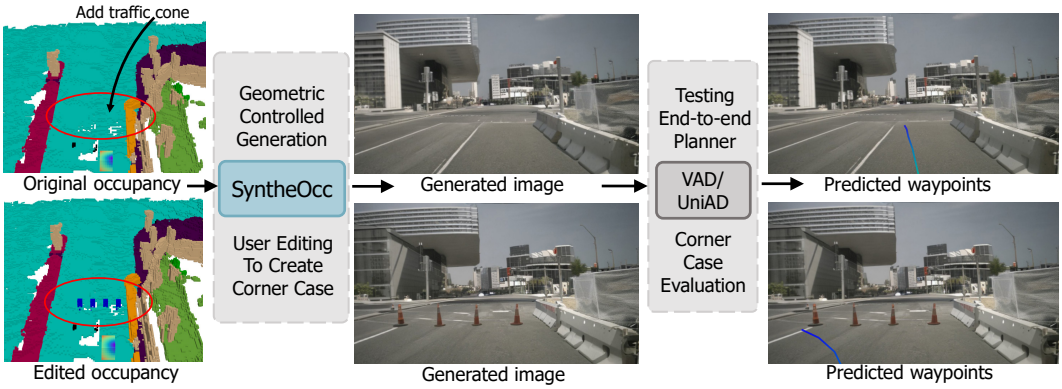

Figure 1: A showcase of application of **SytheOcc**. We enable geometric-controlled generation that conveys the user editing in 3D voxel space to generate realistic street view images. In this case, we create a rare scene that traffic cones block the way. This advancement facilitates the evaluation of autonomous systems, such as the end-to-end planner VAD [9], in simulated corner case scenes.

3D box parameters into text embedding through Fourier mapping that is similar to NeRF [20], and uses cross-attention to learn conditional generation. Although these methods achieve satisfactory results in image generation, their 3D controllability is inherently limited. These approaches are restricted to manipulating the scene in types of 3D boxes and BEV layouts, and hardly adapt to finer geometry control such as editing the shape of objects and scenes. Meanwhile, they usually convert conditional input into 1D embedding that aligns with prompt embedding, which is less effective in 3D-aware generation due to lack of spatial alignment with the generated images. This limitation hinders their utility in downstream applications, such as occupancy prediction and editing scene geometry to create long-tailed scenes, where granular volumetric control is paramount in both tasks.

ControlNet [42] and GLIGEN [14] is another type of prominent method in the field of controllable image generation. These approaches exhibit several desirable attributes in terms of controllability. They leverage conditional images such as semantic masks for control, thereby offering a unified framework to manipulate both foreground and background. However, despite its precise spatial control, ControlNet does not align with our specific requirements. Their conditions of pixel-level images differ fundamentally from what we require in 3D contexts. Our experimental results also find that ControlNet struggles to handle overlapping objects with varying depths (see Fig. 6 (a)), as it only utilizes an ambiguous 2D semantic map as conditional input. As a result, it is non-trivial to extend the ControlNet framework and convey their desirable attributes for 3D conditioning.

To address the above challenges, we propose an innovative representation, 3D semantic multi-plane images (MPIs), which contribute to image generation with finer geometric control. In detail, we employ multi-plane images [44] to represent the occupancy, where each plane represents a slice of semantic label at a specific depth. Our 3D semantic MPIs not only preserve accurate and authentic 3D information, but also keep pixel-wise alignment with the generated images. We additionally introduce the MPI encoder to encode features, and the reweighing methods to ease the training with long-tailed cases. As a collection, our framework enables 3D geometry and semantic control for image generation and further facilitates corner case evaluation as depicted in Fig. 1. Finally, experimental results demonstrate that our synthetic data achieve better recognizability, and are effective in improving the perception model on occupancy prediction. In summary, our contributions include:

- We present **SytheOcc**, a novel image generation framework to attain finer and precise 3D geometric control, thereby unlocking a spectrum of applications such as 3D editing, dataset generation, and long-tailed scene generation.
- Incorporating the proposed 3D semantic MPI, MPI encoder, and reweighing strategy, we deliver a substantial advancement in image quality and recognizability over prior works.
- Our extensive experimental results demonstrate that our synthetic data yields an effective data augmentation in the realm of 3D occupancy prediction.