OccGen: Generative Multi-modal 3D Occupancy Prediction for Autonomous Driving

Guoqing Wang[1], Zhongdao Wang[2], Pin Tang[1], Jilai Zheng[1],
Xiangxuan Ren[1], Bailan Feng[2], and Chao Ma[1]⋆

[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University,
[2] Huawei Noah's Ark Lab
{guoqing.wang,pin.tang,zhengjilai,bunny_renxiangxuan,chaoma}@sjtu.edu.cn
{wangzhongdao,fengbailan}@huawei.com
Project page: https://occgen-ad.github.io/

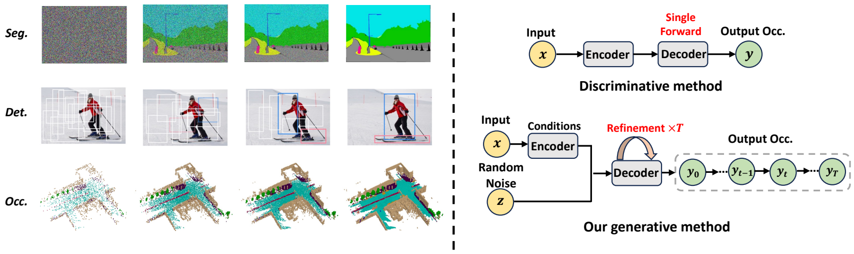
Abstract. Existing 3D semantic occupancy prediction methods typically treat the task as a one-shot 3D voxel-wise segmentation problem, focusing on a single-step mapping between the inputs and occupancy maps, which limits their ability to refine and complete local regions gradually. In this paper, we introduce OccGen, a simple yet powerful generative perception model for 3D semantic occupancy prediction. OccGen adopts a "noise-to-occupancy" generative paradigm, progressively inferring and refining the occupancy map by predicting and eliminating noise originating from a random Gaussian distribution. OccGen consists of two main components: a conditional encoder that is capable of processing multi-modal inputs, and a progressive refinement decoder that applies diffusion denoising using the multi-modal features as conditions. A key insight of this generative pipeline is that the diffusion denoising process is naturally able to model the coarse-to-fine refinement of the dense 3D occupancy map, therefore producing more detailed predictions. Extensive experiments on several occupancy benchmarks demonstrate the effectiveness of the proposed method compared to the state-of-the-art methods. For instance, OccGen relatively enhances the mIoU by 9.5%, 6.3%, and 13.3% on nuScenes-Occupancy dataset under the muli-modal, LiDAR-only, and camera-only settings, respectively. Moreover, as a generative perception model, OccGen exhibits desirable properties that discriminative models cannot achieve, such as providing uncertainty estimates alongside its multiple-step predictions.

Keywords: Occupancy · Generative Model · Diffusion · Multi-modal

1 Introduction

The precise 3D perception of the surrounding environment constitutes the cornerstone of modern autonomous driving systems, as it directly affects downstream tasks such as planning and vehicle control [15, 20]. In recent years, advancements in 3D object detection and segmentation [16, 23, 26–28, 31, 33, 34, 49,

⋆ Corresponding author.



(a) Generative diagram for perception tasks.

(b) Comparison of different pipelines.

Fig. 1: (a) The generative diagram of semantic segmentation (seg.), object detection (det.), and 3D semantic occupancy prediction (occ.). (b) Compared to previous discriminative methods with a single forward evaluation scheme, our OccGen is a generative model that can generate occupancy maps in a coarse-to-fine manner.

50, 57, 59, 60] have propelled the field of 3D perception. However, these methods require either rigid bounding boxes, which oversimplify the object shapes, or Bird's-Eye View (BEV) predictions that involve compromises in projecting 3D scenes onto 2D ground planes. Such methods can significantly impede the ability to accurately perceive structural information along the vertical axis, particularly when dealing with irregular objects.

To address this limitation, 3D semantic occupancy prediction [17,45,47,48,52] has been proposed to assign semantic labels to every spatially occupied region within the perceptive range. Most previous methods for 3D semantic occupancy prediction can be roughly divided into three categories: LiDAR-based [8, 24, 37, 55], vision-based [5, 17, 25, 46, 58], and multi-modal based [51] methods. These methods typically formulate the 3D occupancy prediction as a one-shot voxel-wise segmentation problem with a single forward evaluation scheme. While these works achieve promising results, this perception pipeline faces two critical issues: 1) Discriminative methods primarily focus on learning the mapping between the input-output pairs in a single forward step and neglect the modeling of the underlying occupancy map distribution. 2) Inferring only once is not enough for the model to complete the fine-grained scene well, just like humans need continuous observation to perceive the entire scene fully.

On the other hand, the diffusion model [14,44] has demonstrated its powerful generation capability and has also led to the successful application in numerous discriminative tasks, such as depth estimation [19,41], object detection [6], and segmentation [1, 53, 54]. We observe that the diffusion denoising process is naturally able to model the coarse-to-fine refinement of the dense 3D occupancy map, therefore producing more detailed predictions. Motivated by this, we propose OccGen, a simple yet powerful generative perception model for 3D semantic occupancy. As shown in Fig. 1, OccGen adopts a "noise-to-occupancy" generative paradigm, progressively inferring and eliminating noise originating from a random 3D Gaussian distribution. The proposed OccGen consists of two main components: a conditional encoder and a progressive refinement decoder. The conditional encoder only needs to run once, while the decoder runs multiple times to fulfill progressive refinement. Since the encoder only runs once during the entire inference process, running the decoder step-by-step for diffusion denoising does not introduce significant computational overhead, thereby achieving comparable latency to single forward methods. During the training phase, we obtain a 3D noise map by gradually adding Gaussian noise to the ground truth occupancy. Subsequently, this noise map is fed into the progressive refinement decoder, which utilizes the multi-scale fusion features from the conditional encoder as conditions to generate noise-free predictions. In the inference phase, OccGen progressively generates the occupancy in a coarse-to-fine refinement manner, which is implemented by gradually denoising a 3D Gaussian noise map given the multi-modal condition inputs.

As a generative perception model, OccGen exhibits desirable properties that are not achievable by discriminative models: (1) progressive inference supports trading compute for prediction quality; (2) uncertainty estimation can be readily made alongside the predictions. We evaluate the effectiveness of OccGen on several benchmarks and show promising results compared with the state-of-the-art methods. Notably, OccGen has exhibited performance gains of 9.5%, 6.3%, and 13.3% on mIoU compared with the state-of-the-art method under the mulimodal, LiDAR-only, and camera-only settings on nuScenes-Occupancy.

Our contributions are summarized as follows:

- We introduce OccGen, a simple yet powerful generative framework following the "noise-to-occupancy" paradigm.
- OccGen adopts an efficient design that the encoder only runs once during the entire inference process, and the decoder runs step-by-step for progressive refinement, achieving a comparable latency to single forward methods.
- We extensively validate the proposed OccGen on multiple occupancy benchmarks, demonstrating its remarkable performance and desirable properties compared to previous discriminative methods.

2 Related Work

2.1 3D Semantic Occupancy Prediction

BEV-based methods [16,26–28,31] typically project the 3D scene onto the ground plane, leading to the potential loss of information in the vertical dimension. Compared with BEV representation, 3D semantic occupancy provides a more detailed representation of the environment by explicitly modeling the occupancy status of each voxel in a 3D grid. SSCNet [45] has first introduced the task of semantic scene completion, integrating both geometry and semantics. Subsequent works [8, 24, 37, 55] commonly utilized geometric inputs with explicit depth information. Recently, vision-based occupancy prediction methods [5, 17, 25, 58] have been widely studied due to the cost-effectiveness of cameras. Furthermore, many concurrent works are dedicated to proposing surrounding-view and multi-modal benchmarks for 3D semantic occupancy prediction, contributing to the flourishing of the occupancy community [47, 48, 51, 52]. In this paper, we propose OccGen, a simple yet powerful generative perception framework for 3D multi-modal semantic occupancy that can progressively refine the occupancy in a coarse-to-fine manner.

2.2 Diffusion Model

Diffusion models [14, 43] have been extensively researched due to their powerful generation capability. DDPM [14] proposed a diffusion model where the forward and reverse processes exhibit the Markovian property. DDIM [44] accelerated DDPM [14] by replacing the original diffusion process with non-Markovian chains to enhance sampling speed. On the other hand, conditional diffusion models have also been actively studied. Text-to-image generation models [38] and image-to-image translation models [40] achieved surprising results. Recently, diffusion models for visual perception have attracted widespread attention. Several pioneering works [1, 7, 42, 53, 54] attempted to apply the diffusion model to visual perception tasks, e.g. image segmentation or depth estimation tasks. For all the diffusion models listed above, one or two parameter-heavy convolutional U-Nets [39] are adopted, leading to low efficiency, slow convergence, and sub-optimal performance. DiffusionDet [6] proposed a denoising diffusion process from noisy boxes to object boxes. DDP [19] followed the "noise-to-map" generative paradigm for prediction by progressively removing noise from a random Gaussian distribution, guided by the image. In this work, as illustrated in Fig. 2, we extend the generative diffusion process into the occupancy perception pipeline while maintaining accuracy and efficiency.

3 Method

In this section, we first introduce the preliminaries on 3D semantic occupancy perception and conditional diffusion model. Then, we present the pipeline of the "noise-to-occupancy" and the overall architecture of OccGen. Finally, we show the details of the training and inference process.

3.1 Preliminaries

3D Semantic Occupancy Perception. The objective of 3D semantic occupancy perception is to predict a complete 3D representation of volumetric occupancy and semantic labels for a scene in the surround-view driving scenarios given inputs, such as images and LiDAR points. We utilize LiDAR point cloud $X_p \in \mathbb{R}^{N_L \times (3+d)}$ and multi-view camera images $X_c \in \mathbb{R}^{N_C \times H_C \times W_C \times 3}$ as multi-modal inputs, denoted by $X = \{X_p, X_c\}$. Subsequently, we train a neural network f_θ to generate an occupancy voxel map $Y \in \{c_0, c_1, ..., c_N\}^{H \times W \times Z}$, where each voxel is assigned either an empty label c_0 or occupied by a specific

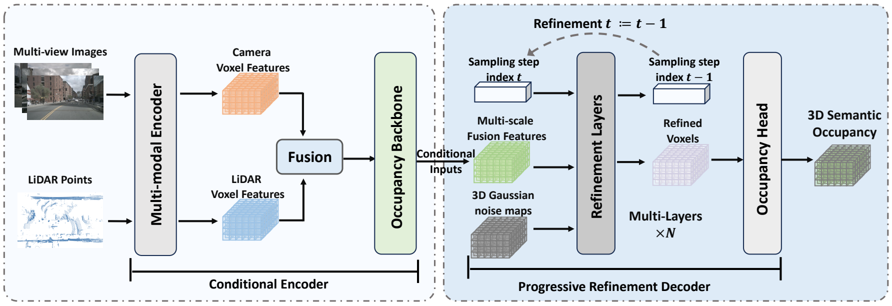


Fig. 2: The overview of the proposed OccGen framework. It has an encoder-decoder structure. The conditional encoder extracts the features from the inputs as the condition. The progressive refinement decoder consists of a stack of refinement layers and an occupancy head, which takes the 3D noise map, sampling step, and conditional multi-scale fusion features as inputs and progressively generates the occupancy prediction.

semantic class from $\{c_1, c_2, ..., c_N\}$. Here, N represents the total number of interested classes, and $\{H, W, D\}$ indicates the volumetric dimensions of the entire scene.

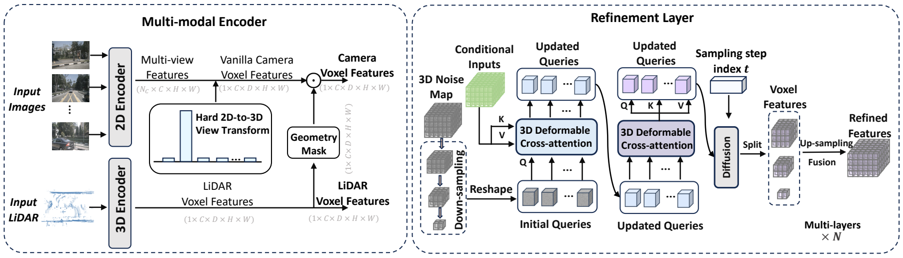
Diffusion Model. The diffusion model is a type of generative model that demonstrates greater potential in the generative domain compared to Generative Adversarial Network (GAN) [11]. It can be divided into two categories: unconditional diffusion models learn an explicit approximation of the data distribution $P(z)$, while conditional diffusion models learn the distribution given a certain condition k, denoted as $p(z|k)$. In the conditional diffusion model, the data distribution is learned by recovering a data sample from Gaussian noise through an iterative denoising process. The forward diffusion process gradually adds noise to the data sample z_0, denoted as:

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

which transforms the z_0 to a latent noisy sample z_t for $t \in \{0, 1, \ldots, T\}$. The constant $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$ and β_s represents the noise schedule. In the training process, the conditional diffusion model $f_\theta\left(\boldsymbol{z}_t, t | \, k\right)$ is trained to predict $\boldsymbol{z}_0$ from $\boldsymbol{z}_t$ under the guidance of condition $\boldsymbol{k}$ by minimizing the training objective function (*i.e.*, l_2 loss). In the inference process, the predicted sample $\boldsymbol{z}_0$ is reconstructed from a random noise $\boldsymbol{z}_T$ with the model f_θ and conditional input $\boldsymbol{k}$ following the denoising process of DDPM [14] or DDIM [44].

3.2 OccGen Framework

We first depict the proposed "noise-to-occupancy" generative paradigm and then introduce the overall architecture. As shown in Fig. 2, OccGen consists of a conditional encoder and a progressive refinement decoder.



(a) The architecture of multi-modal encoder.

(b) The architecture of refinement layer.

Fig. 3: The detailed architectures of (a) multi-modal encoder in conditional encoder and (b) refinement layer in progressive refinement decoder. The multi-modal encoder is a two-stream structure, comprising LiDAR and camera streams. The refinement layer consists of three main components, i.e., 3D deformable cross-attention, self-attention, and time diffusion modules.

Noise-to-Occupancy Generative Paradigm. We regard the 3D semantic occupancy prediction as a generative process, which progressively generates the surrounding 3D environment with detailed geometry and semantics from single or multi-modal inputs. The goal of noise-to-occupancy is to learn an occupancy perception model f_θ which can model the coarse-to-fine refinement of the dense 3D occupancy map through a total of T diffusion steps:

$$Y_T \xrightarrow{f_\theta} Y_{T-1} \xrightarrow{f_\theta} \dots \xrightarrow{f_\theta} Y_0, \tag{2}$$

where the diffusion step $T \rightarrow 0$ represents the coarse-to-fine refinement process from a 3D Gaussian voxel map to the refined occupancy. Thus, the generative occupancy prediction paradigm can be formulated as:

$$\mathit{\Delta} Y_t = f_\theta\left(x, t, Y_{t+1}\right), \quad Y_t = Y_{t+1} \oplus \mathit{\Delta} Y_t, \tag{3}$$

where the model f_θ refines the current prediction occupancy by giving the diffusion step index t and the previous-step prediction occupancy Y_{t+1}, and $\oplus$ is element-wise summation.

Conditional Encoder. The conditional encoder has three main components: a multi-modal encoder, a fusion module, and an occupancy backbone. As shown in Fig. 3(a), the multi-modal encoder is a two-stream structure, comprising of LiDAR and camera streams. For the LiDAR stream, we follow VoxelNet [59] and 3D sparse convolutions [56] to transform raw LiDAR points to LiDAR voxel features. In the camera stream, we utilize the pre-trained 2D backbones [9,13,30] and Feature Pyramid Network (FPN) [29] to extract multi-view image features given multi-view images. We obtain the vanilla camera voxel features through the 2D-to-3D view transformation.

Different from the previous 2D-to-3D view transformation [26, 28, 31, 36] methods that estimate the probabilistic of a set of discrete depths, OccGen proposes a hard 2D-to-3D view transformation to guarantee more accurate depth.

We opt for predicting a one-hot vector for depth, as opposed to utilizing softmax on discrete depth values when lifting each image individually into a frustum of features for each camera. However, obtaining one-hot encoding directly through *argmax* operation is non-differentiable. To address this issue, we utilize Gumbel-Softmax [18] to convert the predicted depth into one-hot encoding.

The previous multi-modal methods for 3D occupancy prediction do not pay much attention to the interaction between multi-modal features. Therefore, we propose a straightforward solution to exploit the geometry-aware correspondence between camera and LiDAR modalities fully. We directly generate a geometry mask by leveraging LiDAR voxel features and then applying it to the vanilla camera voxel features to get the camera voxel features. This feature aggregation strategy effectively bridges the gap between the camera voxel features and the true spatial distribution in the real-world scene. We follow [51] and fuse the camera and LiDAR voxel features using the adaptive fusion module:

$$\begin{aligned} W &= \mathcal{G}_{\mathrm{C}}\left(\left[\mathcal{G}_{\mathrm{C}}\left(F_p\right), \mathcal{G}_{\mathrm{C}}\left(F_c\right)\right]\right), \\ F_m &= \sigma(W) \odot F_p + (1-\sigma(W)) \odot F_c, \end{aligned} \tag{4}$$

where $\mathcal{G}_{\mathrm{C}}$ is the 3D convolution, $[\cdot, \cdot]$ is the concatenation along feature channel. σ and $\odot$ denote the Sigmoid function and element-wise product, respectively. Finally, we fed the multi-modal voxel features F_m into the occupancy backbone to get the multi-scale fusion features for the following progressive refinement decoder. Additional design details and ablations of hard 2D-to-3D view transformation and geometry mask are presented in the supplementary materials.

Progressive Refinement Decoder. The progressive refinement decoder of OccGen consists of a stack of refinement layers and an occupancy head. As illustrated in Fig. 3(b), the refinement layer takes as input the random noise map or the predicted noise map Y_{t+1} from the last step, the current sampling step t, and the multi-scale fusion features F_m. The refinement layer utilizes efficient 3D deformable cross-attention and self-attention to refine the 3D Gaussian noise map. Compared with traditional deformable attention [61] in 2D vision, 3D deformable attention samples the interest points around the reference point in the 3D pixel coordinate system to compute the attention results. Mathematically, 3D deformable attention can be represented by the following general equation:

$$\mathrm{DA}_{\mathrm{3D}}(q, p, F) = \sum_{k=1}^{N} A_k W_k F(p + \Delta p_k), \tag{5}$$

where q and p denote the 3D query and 3D reference point, F represents the flattened 3D voxel features, and k indexes the sampled point from a total of N points around the reference point p. W_k represents the learnable weights for value generation, while A_k corresponds to the learnable attention weight. Δp_k denotes the predicted offset to the reference point p, and $F(p + \Delta p_k)$ signifies the feature at the location $p + \Delta p_k$ extracted through bilinear interpolation. For brevity, we present the formulation for single-head attention only.

Directly operating on the original 3D Gaussian noise map Y_t with high resolution is computationally intensive. Therefore, we first downsample it to obtain smaller multi-scale noise maps $Y_t^i \in \mathbb{R}^{\frac{D}{2^i} \times \frac{H}{2^i} \times \frac{W}{2^i} \times C_i} (i = 1, 2, 3)$. Then, we reshape these downsampled multi-scale noise maps to obtain initial queries. For each initial query q in the multi-scale noise maps Y_t^i, we get the corresponding reference points p on conditional inputs based on their corresponding spatial and level positions. We get the updated queries using 3D deformable cross-attention ($\mathrm{DCA_{3D}}$) by

$$\mathrm{DCA_{3D}}\left(Y_t^i, F_m\right) = \sum_{n \in F_m} \mathrm{DA_{3D}}\left(q, proj(q, n), F_m\right), \tag{6}$$

where n denotes the hit multi-scale features. For each query q, we use $proj$ operation to obtain the reference point on multi-scale fusion features.

After one round of 3D deformable cross-attention, the initial queries gather knowledge from the condition inputs. To further enhance self-completion capability, we utilize the 3D deformable self-attention to update the queries,

$$\mathrm{DSA_{3D}}\left(Y_t^i, Y_t^i\right) = \sum_{n \in Y_t^i} \mathrm{DA_{3D}}\left(q, p, \mathbf{Y}_t^i\right). \tag{7}$$

Then, we split the learned queries into the down-sampled voxel sizes. We further apply a diffusion denoising step on the down-sampled multi-scale noise maps by

$$Y_t^i := \mathrm{Diff}(Y_t^i, \mathrm{ToEmbed}(t)), \tag{8}$$

where $\mathrm{ToEmbed}(\cdot)$ denotes the embedding network that transforms a step index t from scalar into a feature vector. $\mathrm{Diff}(\cdot)$ represents the diffusion module that applies the scale and shift operation along the time embedding. Furthermore, we upsample and project the downsampled voxels to the original 3D noise map and obtain the refined voxel features. Finally, we obtain the 3D semantic occupancy by feeding the refined voxel features to the occupancy head. This process can be performed multiple times to progressively infer and refine the occupancy map by predicting and eliminating noise from a random Gaussian distribution.

3.3 Training

During training, we first construct a denoising diffusion process from the ground truth Y_0 to the 3D Gaussian noise map Y_T and then train the progressive refinement decoder to reverse this process. Detailed information on the training procedure for OccGen is available in the supplementary materials.

Occupancy Corruption. We add Gaussian noise to corrupt the encoded ground truth, obtaining the 3D Gaussian noise map. As shown in Eq. 1, the intensity of corruption noise is controlled by α_t, which follows a monotonically decreasing schedule across different time steps $t \in [0, 1]$. Different noise scheduling strategies, including cosine schedule [35] and linear schedule [14], are compared and discussed in the supplementary materials. We found that the cosine schedule generally yields the best results in 3D semantic occupancy prediction.

Loss Function. The cross-entropy loss $\mathcal{L}_{\text{ce}}$ and lovasz-softmax loss $\mathcal{L}_{\text{ls}}$ [3] are widely used to optimize the networks for semantic segmentation tasks. Following [5,51], we also utilize affinity loss $\mathcal{L}_{\text{scal}}^{\text{geo}}$ and $\mathcal{L}_{\text{scal}}^{\text{sem}}$ to optimize the scene-wise and class-wise metrics (*i.e.*, geometric IoU and semantic mIoU). Additionally, the depth loss $\mathcal{L}_{\text{d}}$ [26] is used to optimize the predicted depth. Therefore, the overall loss function can be derived as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{ls}} + \mathcal{L}_{\text{scal}}^{\text{geo}} + \mathcal{L}_{\text{scal}}^{\text{sem}} + \mathcal{L}_{\text{d}}. \tag{9}$$

3.4 Inference

Given multi-scale fusion features as conditional inputs, OccGen samples a random noise map from a 3D Gaussian distribution and produces the occupancy in a coarse-to-fine manner. The inference procedure for OccGen is provided in supplementary materials.

Sampling Rule. Following [19], we choose the DDIM strategy [44] for the sampling. In each sampling step, the random noise map or predicted noise map from the last step and the conditional multi-scale fusion features are sent to the progressive refinement decoder for occupancy prediction. After obtaining the predicted result of the current step, we compute the refined noise map for the next step using the reparameterization trick. Following [6,19], we use the asymmetric time intervals (controlled by a hyper-parameter td) during the inference stage. We empirically set $td = 1$ in our method.

Progressively Inference. According to the feature that the diffusion model can generate the distribution step by step, we can perform progressive inference to get fine-grained occupancy in a coarse-to-fine manner. Moreover, OccGen has a natural awareness of the prediction uncertainty. As a comparison, previous one-shot approaches for 3D semantic occupancy [25,51,52,58] can only output a certain occupancy during the inference stage, and are unable to assess the reliability and uncertainty of model predictions.

4 Experiments

4.1 Experimental Setup

Dataset and Metrics. We evaluate our proposed OccGen on two benchmarks, i.e., nuScenes-Occupancy [51] and SemanticKITTI [2]. The nuScenes-Occupancy extends the nuScenes [4] to provide dense annotations on keyframes for 3D multi-modal semantic occupancy prediction. It covers 700 and 150 driving scenes in the training and validation sets of nuScenes. SemanticKITTI [2] contains 22 sequences including monocular images, LiDAR points, point cloud segmentation labels, and semantic scene completion annotations. We follow previous works [25,

Table 1: Semantic occupancy prediction results on nuScenes-Occupancy validation set. The C, D, L, M denotes **camera, depth, LiDAR** and **multi-modal**. For **Surround=✓**, the method directly predicts surrounding semantic occupancy with 360-degree inputs. Otherwise, the method produces the results of each camera view and then concatenates them as surrounding outputs. Best camera-only, LiDAR-only, and multi-modal results are marked as **red**, **blue**, and **black**, respectively.

Method	Input	Surround	IoU	mIoU	barrier ■	bicycle ■	bus ■	car ■	const. veh. ■	motorcycle ■	pedestrian ■	traffic cone ■	trailer ■	truck ■	drive. suf. ■	other flat ■	sidewalk ■	terrain ■	manmade ■	vegetation ■
MonoScene [5]	C	✗	18.4	6.9	7.1	3.9	9.3	7.2	5.6	3.0	5.9	4.4	4.9	4.2	14.9	6.3	7.9	7.4	10.0	7.6
TPVFormer [17]	C	✓	15.3	7.8	9.3	4.1	11.3	10.1	5.2	4.3	5.9	5.3	6.8	6.5	13.6	9.0	8.3	8.0	9.2	8.2
3DSketch [8]	C&D	✗	25.6	10.7	12.0	5.1	10.7	12.4	6.5	4.0	5.0	6.3	8.0	7.2	21.8	14.8	13.0	11.8	12.0	21.2
AICNet [24]	C&D	✗	23.8	10.6	11.5	4.0	11.8	12.3	5.1	3.8	6.2	6.0	8.2	7.5	24.1	13.0	12.8	11.5	11.6	20.2
LMSCNet [37]	L	✓	27.3	11.5	12.4	4.2	12.8	12.1	6.2	4.7	6.2	6.3	8.8	7.2	24.2	12.3	16.6	14.1	13.9	22.2
JS3C-Net [55]	L	✓	30.2	12.5	14.2	3.4	13.6	12.0	7.2	4.3	7.3	6.8	9.2	9.1	27.9	15.3	14.9	16.2	14.0	**24.9**
C-OpenOccupancy [51]	C	✓	19.3	10.3	9.9	6.8	11.2	11.5	6.3	8.4	8.6	4.3	4.2	9.9	22.0	15.8	14.1	13.5	7.3	10.2
L-OpenOccupancy [51]	L	✓	30.8	11.7	12.2	4.2	11.0	12.2	8.3	4.4	8.7	4.0	8.4	10.3	23.5	16.0	14.9	15.7	15.0	17.9
OpenOccupancy [51]	M	✓	29.1	15.1	14.3	12.0	15.2	14.9	13.7	15.0	13.1	9.0	10.0	14.5	23.2	17.5	16.1	17.2	15.3	19.5
C-CONet [51]	C	✓	20.1	12.8	13.2	8.1	**15.4**	17.2	6.3	11.2	10.0	8.3	4.7	12.1	31.4	18.8	18.7	16.3	4.8	8.2
L-CONet [51]	L	✓	30.9	15.8	17.5	**5.2**	13.3	18.1	7.8	5.4	9.6	5.6	13.2	13.6	34.9	21.5	22.4	21.7	19.2	23.5
CONet [51]	M	✓	29.5	20.1	23.3	13.3	21.2	24.3	**15.3**	15.9	18.0	13.3	15.3	20.7	33.2	21.0	22.5	21.5	19.6	23.2
C-OccGen	C	✓	**23.4**	**14.5**	**15.5**	**9.1**	15.3	**19.2**	**7.3**	**11.3**	**11.8**	**8.9**	**5.9**	**13.7**	**34.8**	**22.0**	**21.8**	**19.5**	**6.0**	**9.9**
L-OccGen	L	✓	**31.6**	**16.8**	**18.8**	5.1	**14.8**	**19.6**	**7.0**	**7.7**	**11.5**	**6.7**	**13.9**	**14.6**	**36.4**	**22.1**	**22.8**	**22.3**	**20.6**	24.5
OccGen	M	✓	**30.3**	**22.0**	**24.9**	**16.4**	**22.5**	**26.1**	14.0	**20.1**	**21.6**	**14.6**	**17.4**	**21.9**	**35.8**	**24.5**	**24.7**	**24.0**	**20.5**	**23.5**

51, 58] to report the Intersection of Union (IoU) as the geometric metric and the mean Intersection over Union (mIoU) of each class as the semantic metric.

Implementation Details. We follow the same experiment settings of [51, 58] to make a fair comparison with previous methods [5, 25, 51, 58] on both nuScenes-Occupancy and SemanticKITTI. We stack six refinement layers with 3D deformable attention for the progressive refinement decoder. For training, we leverage the AdamW [22] optimizer with a weight decay of 0.01 and an initial learning rate of $2e-4$. We adopt the cosine learning rate scheduler with linear warming up in the first 500 iterations, and a similar augmentation strategy as BEVDet [16]. The models are trained for 24 epochs with a batch size of 8 on 8 V100 GPUs. More implementation details are listed in supplementary materials.

4.2 Comparison with the state-of-the-art

Results on nuScenes-Occupancy. As shown in Tab. 1, we report the quantitative comparison of existing LiDAR-based, camera-based, and multi-modal methods on nuScenes-Occupancy. Compared with the current SOTA method CONet [51], OccGen achieves a remarkable boost of 1.7%, 1.0%, and 1.9% mIoU for camera-only, LiDAR-only, and multi-modal settings, respectively. This demonstrates the effectiveness of OccGen for semantic occupancy prediction. We also note that OccGen consistently delivers the best IoU results across almost all categories, which indicates that our method can better complete the scenes due to our coarse-to-fine generation property. It is also worth noting that OccGen with multi-modal inputs can improve camera-only and LiDAR-only by 7.5%

Table 2: Semantic Scene Completion results on SemanticKITTI [2] validation set. † denotes the results provided by MonoScene [5].

Method	IoU	mIoU	road. (%) ■	sidewalk. (%) ■	parking. (%) ■	otherground. (%) ■	building. (%) ■	car. (%) ■	truck. (%) ■	bicycle. (%) ■	motorcycle. (%) ■	othervehicle. (%) ■	vegetation. (%) ■	trunk. (%) ■	terrain. (%) ■	person. (%) ■	bicyclist. (%) ■	motorcyclist. (%) ■	fence. (%) ■	pole. (%) ■	trafficsign. (%) ■
LMSCNet† [37]	28.61	6.70	40.68	18.22	4.38	0.00	10.31	18.33	0.00	0.00	0.00	0.00	13.66	0.02	20.54	0.00	0.00	0.00	1.21	0.00	0.00
AICNet† [24]	29.59	8.31	43.55	20.55	11.97	0.07	12.94	14.71	4.53	0.00	0.00	0.00	15.37	2.90	28.71	0.00	0.00	0.00	2.52	0.06	0.00
JS3C-Net† [55]	38.98	10.31	50.49	23.74	11.94	0.07	15.03	24.65	4.41	0.00	0.00	6.15	18.11	4.33	26.86	0.67	0.27	0.00	3.94	3.77	1.45
MonoScene [5]	37.12	11.50	57.47	27.05	15.72	**0.87**	14.24	23.55	7.83	0.20	0.77	3.59	18.12	2.57	30.76	1.79	1.03	0.00	6.39	4.11	2.48
TPVFormer [17]	35.61	11.36	56.50	25.87	20.60	0.85	13.88	23.81	8.08	0.36	0.05	4.35	16.92	2.26	30.38	0.51	0.89	0.00	5.94	3.14	1.52
VoxFormer [25]	44.02	12.35	54.76	26.35	15.50	0.70	17.65	25.79	5.63	0.59	0.51	3.77	24.39	**5.08**	29.96	1.78	3.32	0.00	**7.64**	7.11	4.18
OccFormer [58]	36.50	13.46	58.85	26.88	19.61	0.31	14.40	25.09	**25.53**	0.81	1.19	8.52	19.63	3.93	**32.62**	2.78	2.82	0.00	5.61	4.26	2.86
Symphonize [21]	41.44	13.44	55.78	26.77	14.57	0.19	**18.76**	**27.23**	15.99	**1.44**	2.28	9.52	**24.50**	4.32	28.49	3.19	**8.09**	0.00	6.18	**8.99**	**5.39**
OccGen (ours)	36.87	**13.74**	**61.28**	**28.30**	20.42	0.43	14.49	26.83	15.49	**1.60**	**2.53**	**12.83**	20.04	3.94	32.44	**3.20**	3.37	0.00	6.94	4.11	2.77

and 5.2% mIoU, which demonstrates the effectiveness of the camera modality in capturing small objects (e.g., bicycle, pedestrian, motorcycle, traffic cone) and LiDAR modality on large objects structured regions (e.g., drivable surface, sidewalk, vegetation). This lays a solid foundation for us to further explore how to improve the role of images during fusion.

Results on SemanticKITTI. We also compare OccGen with the state-of-the-art vision-based works [21,25,58] on SemanticKITTI. For a fair comparison under the camera-only setting, we removed the LiDAR stream and fusion module from the conditional encoder. As shown in Tab. 2, we can see that OccGen achieves the highest mIoU compared with all existing competitors. Compared with the state-of-the-art OccFormer [58], our proposed method has an improvement of 0.3% mIoU, demonstrating the effectiveness of OccGen for semantic scene completion. We also notice that the transformer-based methods [10,21,25,58] achieve higher performance than other previous methods. This reveals the superior capability of transformer-based structure in representation learning.

4.3 Ablation Study

Overall Architecture. The ablation results on the conditional encoder and progressive refinement decoder are shown in Tab. 3. Both the conditional encoder and progressive refinement decoder can achieve performance improvement. We also notice that "with proposed decoder" has a higher performance than "with proposed encoder", demonstrating the effectiveness of our generative pipeline.

Conditional Encoder. We also conduct the ablations on the detailed components of the conditional encoder. From Tab. 4, We also observe that the two solutions in the conditional encoder have both achieved promising performance. The reason is that the accurate depth estimation and geometry guidance can keep the fine-grained spatial structures. This effectively limits the impact of disruptive information from the images, leading to notable performance enhancements.

Table 3: Ablations on the conditional encoder and progressive refinement decoder on nuScenes-Occupancy under multi-modal setting. (a), (b) and (c) denote our baseline, baseline "with encoder" and "with decoder", respectively.

	Encoder	Decoder	IoU	mIoU
(a)	-	-	28.1	20.4
(b)	✓	-	28.6	20.7
(c)	-	✓	30.1	21.6
(d)	✓	✓	30.3	22.0

Table 4: Ablations on the multi-modal encoder on nuScenes-Occupancy under multi-modal setting.'**Hard LSS**" and "**Geo. Mask**" denotes the hard 2D-to-3D view transformation and geometry mask modules in multi-modal encoder.

Hard LSS	Geo. Mask	IoU	mIoU
-	-	29.8	21.4
✓	-	30.2	21.5
-	✓	30.3	21.6
✓	✓	30.3	22.0

Progressive Refinement Decoder. We conduct the ablations on the detailed components of the progressive refinement decoder. From Tab. 5 (a) and (b), it is evident that both 3D deformable cross- and self-attention lead to noticeable improvements in results. Compared to self-attention, cross-attention has a greater impact on performance, which is intuitive: learning knowledge from conditional inputs is always more comprehensive. Additionally, we also observed that the order of DCA and DSA in the decoder has a certain impact on the results. We also see that removing the temporal diffusion process leads to a decrease in results from Tab. 5(c).

4.4 Further Discussion

The desirable properties of OccGen compared with the previous discriminative occupancy methods in a single-forward process are shown in Fig. 4 and Fig. 6. OccGen provides the flexibility to balance computational cost against prediction quality in a coarse-to-fine manner. Additionally, the stochastic sampling process enables the computation of voxel-wise uncertainty in the prediction.

Progressive Refinement. We evaluate OccGen with one, three, and six refinement layers by increasing their sampling steps from one to ten. The results are presented in Fig 5. It can be seen that OccGen can continuously improve its performance by using more sampling steps. For instance, OccGen with six refinement layers shows an increase from 21.7% mIoU (first step) to 22.0% mIoU (third step), and we visualize the inference results of different steps in Fig. 4. In comparison to the previous single-step discriminative method, OccGen has the flexibility to balance computational cost against accuracy. This means our method can be adapted to different trade-offs between speed and accuracy under various scenarios without the need to retrain the network.

Efficiency vs. Accuracy. We report the results of IoU and mIoU to represent the accuracy of different methods and latency(ms) to represent the efficiency of the models. The results are shown in Tab. 6. Compared with the representative discriminative methods, OccGen achieves better results than state-of-the-art CONet [51] when using only one sampling step, with comparable latency on the camera-only, and multi-modal settings. When adopting two sampling steps, the

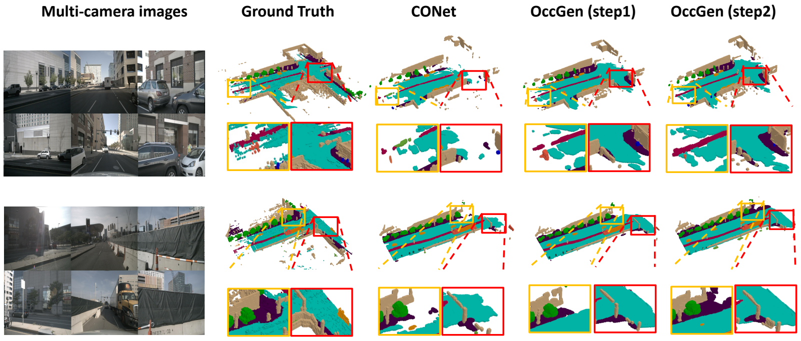


Fig. 4: Qualitative results of occupancy predictions on nuScenes-Occupancy. The leftmost column shows the input surrounding images, and the following four columns visualize the 3D semantic occupancy results from the ground truth, CONet [51], OccGen(first step), and OccGen(second step). The regions highlighted by rectangles indicate that these areas have obvious differences (better viewed when zoomed in).

Table 5: Ablations on the designed components of progressive refinement decoder on nuScenes-Occupancy under multi-modal setting.

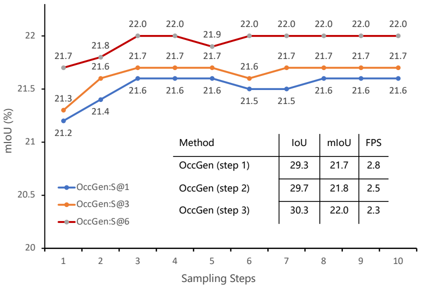
	Method	IoU	mIoU
(a)	w/o DSA	30.1	21.4
	w/o DCA	29.7	21.2
	w/o DCA and DSA	29.1	20.7
(b)	DSA + DCA	29.4	21.6
	DCA + DSA	30.3	22.0
(c)	w/o Diffusion	29.3	21.7
	OccGen	30.3	22.0

Table 6: The latency (ms) and performance (%) of baselines and OccGen on nuScenes-Occupancy under camera-only and multi-modal settings.

Models	Latency(ms)	IoU	mIoU
C-Baseline [51]	172.4	19.3	10.3
C-CONet [51]	285.7	20.1	12.8
C-OccGen(step1)	294.1	23.0	14.2
C-OccGen(step2)	312.5	23.3	14.4
Baseline [51]	243.9	29.1	15.1
CONet [51]	344.8	29.5	20.1
OccGen(step1)	357.1	29.3	21.7
OccGen(step2)	400.0	29.7	21.8

performance is further boosted to 21.8% and 14.4% on the multi-modal and camera-only benchmarks, at a loss of 20 $\sim$ 50 ms. These results show that OccGen can progressively refine the output occupancy multiple times with reasonable time cost.

Uncertainty Awareness. In addition to the performance gains, the proposed OccGen can naturally provide uncertainty estimates. In the multi-step sampling process, we can simply count the voxels where the predicted result of each step differs from the result of the previous step, thereby obtaining an uncertainty occupancy result. We can see from Fig. 6 that the areas with high uncertainty in the uncertainty map often align with those in the error map, which indicates incorrect prediction regions. In comparison, OccGen offers a straightforward and inherently capable approach, whereas previous methods [12, 32] require complicated modeling such as Bayesian networks.



Method	IoU	mIoU	FPS
OccGen (step 1)	29.3	21.7	2.8
OccGen (step 2)	29.7	21.8	2.5
OccGen (step 3)	30.3	22.0	2.3

Fig. 5: The results of multiple inferences on nuScenese-Occupancy under the multi-modal setting.

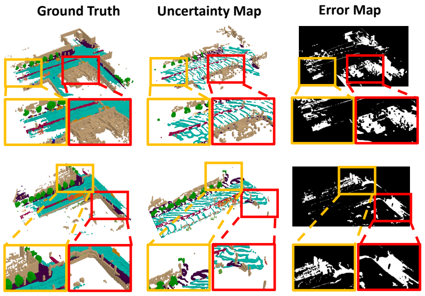


Fig. 6: The visualization of uncertainty map and error map on nuScenese-Occupancy under the multi-modal setting.

4.5 Qualitative Results

In Fig. 4, we visualize the predicted results of 3D semantic occupancy on nuScenes-Occupancy from CONet [51] and our proposed OccGen. Compared with CONet, our method can better understand the scene-level semantic layout and perform local region completion. It is obvious that the regions of "drivable surface" and "sidewalk" predicted by our OccGen have higher continuity and integrity, and can effectively reduce a large number of hole areas compared with CONet. One more interesting observation is that due to the ground truth being initially constructed based on sparse LiDAR data, the shape of voxels in space is not very well-defined, especially in the drivable area. However, both CONet [51] and OccGen yields smoother predictions for these occupancy results.

5 Conclusion

We propose OccGen, a simple yet powerful generative perception model for 3D semantic occupancy prediction. OccGen adopts a "noise-to-occupancy" generative paradigm, progressively inferring and refining the occupancy map from a random Gaussian distribution. OccGen consists of two main components: a conditional encoder that processes the multi-modal inputs and a progressive refinement decoder that produces fine-grained occupancy in a coarse-to-fine manner. OccGen has achieved state-of-the-art performance on several occupancy benchmarks and shown desirable properties that discriminative models cannot achieve, such as progressive inference and uncertainty estimates. Currently, the latency of our OccGen is comparable to the previous state-of-the-art methods and has not achieved a significant speed advantage. Next, we will explore a more lightweight generative architecture for 3D semantic occupancy prediction.

Acknowledgements. This work was supported by NSFC (62322113, 62376156), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Fundamental Research Funds for the Central Universities.

References

1. Amit, T., Nachmani, E., Shaharbany, T., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
2. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: ICCV. pp. 9297–9307 (2019)
3. Berman, M., Triki, A.R., Blaschko, M.B.: The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: CVPR. pp. 4413–4421 (2018)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR. pp. 11621–11631 (2020)
5. Cao, A.Q., de Charette, R.: Monoscene: Monocular 3d semantic scene completion. In: CVPR. pp. 3991–4001 (2022)
6. Chen, S., Sun, P., Song, Y., Luo, P.: Diffusiondet: Diffusion model for object detection. In: ICCV. pp. 19830–19843 (2023)
7. Chen, T., Li, L., Saxena, S., Hinton, G., Fleet, D.J.: A generalist framework for panoptic segmentation of images and videos. arXiv preprint arXiv:2210.06366 (2022)
8. Chen, X., Lin, K.Y., Qian, C., Zeng, G., Li, H.: 3d sketch-aware semantic scene completion via semi-supervised structure prior. In: CVPR. pp. 4193–4202 (2020)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Everingham, M., Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2009)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS. vol. 27 (2014)
12. Harakeh, A., Smart, M., Waslander, S.L.: Bayesod: A bayesian approach for uncertainty estimation in deep object detectors. In: ICRA. pp. 87–93. IEEE (2020)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33, pp. 6840–6851 (2020)
15. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: CVPR. pp. 17853–17862 (2023)
16. Huang, J., Huang, G., Zhu, Z., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
17. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3d semantic occupancy prediction. In: CVPR. pp. 9223–9232 (2023)
18. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016)
19. Ji, Y., Chen, Z., Xie, E., Hong, L., Liu, X., Liu, Z., Lu, T., Li, Z., Luo, P.: Ddp: Diffusion model for dense visual prediction. In: ICCV. pp. 21741–21752 (2023)
20. Jia, X., Gao, Y., Chen, L., Yan, J., Liu, P.L., Li, H.: Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In: CVPR. pp. 7953–7963 (2023)

21. Jiang, H., Cheng, T., Gao, N., Zhang, H., Liu, W., Wang, X.: Symphonize 3d semantic scene completion with contextual instance queries. In: CVPR. pp. 20258–20267 (2024)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: CVPR. pp. 12697–12705 (2019)
24. Li, J., Han, K., Wang, P., Liu, Y., Yuan, X.: Anisotropic convolutional networks for 3d semantic scene completion. In: CVPR. pp. 3351–3359 (2020)
25. Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., Feng, C., Anandkumar, A.: Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In: CVPR. pp. 9087–9098 (2023)
26. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. arXiv preprint arXiv:2206.10092 (2022)
27. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: ECCV. pp. 1–18. Springer (2022)
28. Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., Tang, Z.: Bevfusion: A simple and robust lidar-camera fusion framework. In: NeurIPS (2022)
29. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)
30. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: CVPR. pp. 10012–10022 (2021)
31. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: ICRA (2023)
32. Loquercio, A., Segu, M., Scaramuzza, D.: A general framework for uncertainty estimation in deep learning. IEEE Robotics and Automation Letters **5**(2), 3153–3160 (2020)
33. Lu, H., Tang, J., Xu, X., Cao, X., Zhang, Y., Wang, G., Du, D., Chen, H., Chen, Y.: Scaling multi-camera 3d object detection through weak-to-strong eliciting. arXiv preprint arXiv:2404.06700 (2024)
34. Lu, H., Zhang, Y., Lian, Q., Du, D., Chen, Y.: Towards generalizable multi-camera 3d object detection via perspective debiasing. arXiv preprint arXiv:2310.11346 (2023)
35. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML. pp. 8162–8171. PMLR (2021)
36. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: ECCV. pp. 194–210. Springer (2020)
37. Roldao, L., de Charette, R., Verroust-Blondet, A.: Lmscnet: Lightweight multiscale 3d semantic completion. In: 3DV (2020)
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
39. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
40. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: SIGGRAPH. pp. 1–10 (2022)

41. Saxena, S., Kar, A., Norouzi, M., Fleet, D.J.: Monocular depth estimation using diffusion models. arXiv preprint arXiv:2302.14816 (2023)
42. Saxena, S., Kar, A., Norouzi, M., Fleet, D.J.: Monocular depth estimation using diffusion models. arXiv preprint arXiv:2302.14816 (2023)
43. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML. pp. 2256–2265. PMLR (2015)
44. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
45. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: CVPR. pp. 1746–1754 (2017)
46. Tang, P., Wang, Z., Wang, G., Zheng, J., Ren, X., Feng, B., Ma, C.: Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. arXiv preprint arXiv:2404.09502 (2024)
47. Tian, X., Jiang, T., Yun, L., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. arXiv preprint arXiv:2304.14365 (2023)
48. Tong, W., Sima, C., Wang, T., Chen, L., Wu, S., Deng, H., Gu, Y., Lu, L., Luo, P., Lin, D., et al.: Scene as occupancy. In: ICCV. pp. 8406–8415 (2023)
49. Vora, S., Lang, A.H., Helou, B., Beijbom, O.: Pointpainting: Sequential fusion for 3d object detection. In: CVPR. pp. 4604–4612 (2020)
50. Wang, C., Ma, C., Zhu, M., Yang, X.: Pointaugmenting: Cross-modal augmentation for 3d object detection. In: CVPR. pp. 11794–11803 (2021)
51. Wang, X., Zhu, Z., Xu, W., Zhang, Y., Wei, Y., Chi, X., Ye, Y., Du, D., Lu, J., Wang, X.: Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In: ICCV. pp. 17850–17859 (2023)
52. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In: ICCV. pp. 21729–21740 (2023)
53. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. In: MIDL. pp. 1336–1348 (2022)
54. Wu, J., Fang, H., Zhang, Y., Yang, Y., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. arXiv preprint arXiv:2211.00611 (2022)
55. Yan, X., Gao, J., Li, J., Zhang, R., Li, Z., Huang, R., Cui, S.: Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In: AAAI. vol. 35, pp. 3101–3109 (2021)
56. Yan, Y., Mao, Y., Li, B.: SECOND: Sparsely embedded convolutional detection. Sensors (2018)
57. Zhang, Y., Zhou, Z., David, P., Yue, X., Xi, Z., Gong, B., Foroosh, H.: Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In: CVPR. pp. 9601–9610 (2020)
58. Zhang, Y., Zhu, Z., Du, D.: Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In: ICCV. pp. 9433–9443 (2023)
59. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: CVPR. pp. 4490–4499 (2018)
60. Zhu, X., Zhou, H., Wang, T., Hong, F., Ma, Y., Li, W., Li, H., Lin, D.: Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In: CVPR. pp. 9939–9948 (2021)
61. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2021), https://openreview.net/forum?id=gZ9hCDWe6ke