Occupancy as Set of Points

Yiang Shi[1,⋆], Tianheng Cheng[1,⋆], Qian Zhang[2],
Wenyu Liu[1], and Xinggang Wang[1,✉]

[1] School of EIC, Huazhong University of Science & Technology
[2] Horizon Robotics

Abstract. In this paper, we explore a novel point representation for 3D occupancy prediction from multi-view images, which is named *Occupancy as Set of Points*. Existing camera-based methods tend to exploit dense volume-based representation to predict the occupancy of the whole scene, making it hard to focus on the special areas or areas out of the perception range. In comparison, we present the *Points of Interest* (PoIs) to represent the scene and propose OSP, a novel framework for point-based 3D occupancy prediction. Owing to the inherent flexibility of the point-based representation, OSP achieves strong performance compared with existing methods and excels in terms of training and inference adaptability. It extends beyond traditional perception boundaries and can be seamlessly integrated with volume-based methods to significantly enhance their effectiveness. Experiments on the Occ3D-nuScenes occupancy benchmark show that OSP has strong performance and flexibility. Code and models are available at https://github.com/hustvl/osp.

Keywords: 3D Occupancy Prediction · Autonomous Vehicles · Multiview 3D Perception

1 Introduction

Holistic 3D scene understanding is crucial for autonomous driving systems, directly affecting the efficiency and accuracy of subsequent tasks. Considering the cost-effectiveness and ease of deployment of cameras compared to other sensors, developing visual-based methods for 3D scene understanding has become a significant and widely researched challenge.

To tackle this challenge, 3D Semantic Scene Completion (SSC) [22] has been proposed and widely studied to jointly infer the geometry and semantics information of the scene from limited observations. The SSC task requires the model to accurately predict the visible locations and complete the information for the invisible locations. Recently, Occ3D [25] introduces a new task definition called 3D occupancy prediction. The main difference between this task and SSC is that 3D occupancy prediction only focuses on the visible areas and is tailored for dynamic scenes.

⋆ Equal contribution.
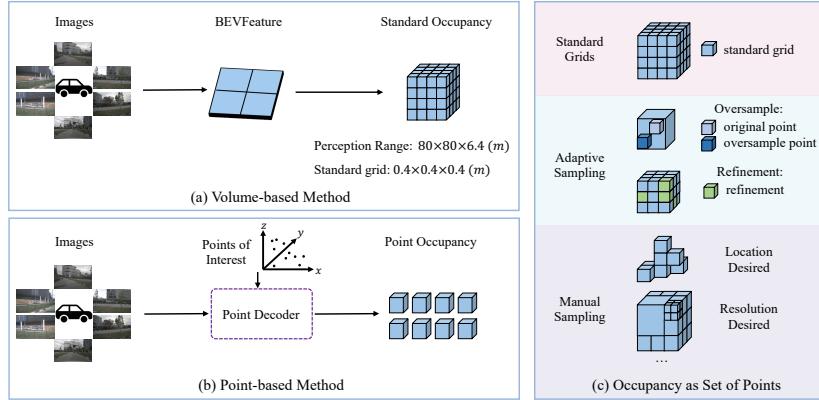✉ Corresponding author: Xinggang Wang (xgwang@hust.edu.cn).

Fig. 1: Comparison between volume-based methods and our method. The volume-based methods, represented by BEVFormer, infers every region within the scene and gets standard occupancy as shown in (a). Our method uses a point-based decoder as shown in (b). Thus it infers the **Points of Interest** including standard, adaptively sampled, and manually sampled grids as shown in (c).

Existing 3D occupancy prediction methods are mostly based on dense BEV methods, *e.g.*, BEVFormer [15], BEVDet [7]. These methods integrate a BEV encoder with an occupancy head to generate the output and enhance BEV perception capabilities for better results. However, they share some common drawbacks. (1) *Uniform Sampling*: BEV-based methods fail to differentiate between different areas within the same scene, treating them equally. This leads to coarse sampling and hinders dynamic or multi-resolution sampling capabilities. (2) *Limited Inference Flexibility*: During inference, these methods can only process the entire scene at once. They lack the ability to infer different parts of the scene based on varying downstream tasks or specific practical needs.

Those limitations highlight the need for more flexible 3D occupancy prediction methods that can handle complex scenes while adapting to different inference requirements. In this paper, we propose a novel point-based representation for 3D occupancy prediction. Instead of dividing the scene into uniform grids applied by existing volume-based methods, we propose *Points of Interest* (PoIs) to view the scene as a collection of points that help in flexibly sampling the scene during both training and inference stages. Fig. 1 compares volume-based and point-based representations. Compared to volume-based representations, our point-based representation has the following advantages: (1) it can accept inputs of any scale and position to make occupancy predictions, including manually designed and adaptively designed input, offering flexibility; (2) it can pay extra attention to certain areas rather than treating all areas equally, enhancing the model's perceptual capabilities.

We introduce Occupancy as Set of Points (OSP), a novel and flexible point-based framework, which is built upon the foundational concept of Points of

Interest (PoIs). OSP excels in 3D occupancy prediction and is composed of an image backbone, a 3D positioning encoder, and a decoder, as illustrated in Fig. 2. Central to our methodology is the innovative use of PoIs, which we have categorized into three distinct types to meet diverse needs, thereby significantly enhancing various aspects of our model's performance. Detailed descriptions of these PoIs will be provided in Sec. 3. Notably, PoIs can be designed as needed beyond the three types initially proposed by us.

Our method stands out for its strong performance and flexibility. The flexibility enables it to process any arbitrary local scene without necessitating retraining. Additionally, our method can serve as an augmentative plugin module for existing volume-based methods by adaptively resampling areas of low confidence to yield more accurate occupancy predictions. It is also adept at predicting areas beyond the scene. The key contributions of our approach can be summarized as follows:

– A novel point-based occupancy representation, established by interacting point queries with 2D image features, enables a comprehensive understanding of 3D scenes.
– A flexible framework that allows for inference at any area of interest without retraining or sacrificing accuracy and predicts areas beyond the scene.
– A plugin module that enhances the performance of the volume-based baseline significantly.
– OSP has obtained strong experiment results, *i.e.*, 39.4 mIoU in the 3D occupancy prediction task on the Occ3D-nuScenes benchmark.

2   Related Work

2.1   3D Occupancy Prediction

3D occupancy prediction, a concept recently defined by Occ3D [25], exhibits notable parallels with Occupancy Grid Mapping (OGM) [20, 24, 27] used in robotics. This task aims to predict the state of each voxel grid in a scene based on a series of sensor inputs. Occ3D establishes two benchmarks leveraging the Waymo Open Dataset [23] and the nuScenes Dataset [2] to facilitate this. In vision-based 3D Occupancy prediction, Occ3D implements camera visibility estimation and creates visibility masks to ensure evaluations are confined to visible areas. It also evaluates various SSC methodologies on its benchmarks, including MonoScene [4], TPVFormer [9], BEVDet [7], OccFormer [32], and BEV-Former [15].

2.2   3D Semantic Scene Completion

Scene Semantic Completion (SSC) represents a task closely associated with 3D occupancy prediction. The concept of SSC is initially presented in SSC-Net [22], with a focus on predicting the comprehensive semantic information of a scene based on its partially visible regions. Over recent years, the study

of SSC has expanded significantly, particularly in the context of small indoor scenes [3, 5, 11–13, 16, 30, 31]. In recent times, the study of Scene Semantic Completion (SSC) for expansive outdoor environments has gained momentum, particularly following the introduction of the SemanticKITTI dataset [1]. Notably, MonoScene [4] emerges as the first method to apply monocular pure vision-based SSC. In a parallel advancement, OccDepth [19] enhances 2D to 3D feature transformation by incorporating depth data from stereo input. TPVFormer [9] argues against the limitations of single-plane modeling in capturing intricate details, hence it adopts a tri-perspective view (TPV) approach, combining a Bird's Eye View (BEV) with two additional vertical planes. Additionally, Symphonies [10] highlights the significance of instance representation in SSC tasks. While SSC methods can be directly applied to 3D occupancy prediction, two primary distinctions exist: (1) SSC primarily aims to infer the occupancy of non-visible areas based on visible regions, in contrast to 3D occupancy prediction which focuses on visible areas; (2) SSC methods usually target static scenes, whereas 3D occupancy prediction methods are often designed to handle dynamic scenes.

Most existing volume-based SSC and 3D occupancy prediction methods are characterized by their dense nature, encompassing inputs and outputs that span the entire scene. Consider the BEVFormer baseline as an example: it segments the scene into uniform BEV grids, failing to distinguish between grids in varying areas. This uniformity restricts the ability of volume-based methods like BEVFormer to sample areas of interest for better performance during training. Besides, if we want to focus on a specific area in the inference stage, volume-based methods are limited to infer the entire scene and then perform post-processing, inevitably leading to increased and unnecessary costs. Moreover, as scene size and voxel resolution increase, the computational demands skyrocket exponentially. In stark contrast, our point-based model introduces much-needed flexibility by focusing on PoIs. Our method facilitates direct inference in specific areas, eliminating the need for post-processing and avoiding additional computational burdens.

A point-related SSC method is PointOcc [34], a point cloud-based SSC prediction method using three complementary view planes for efficient point cloud feature modeling and an efficient 2D backbone for processing to reduce computational load, while our method focuses on the flexibility of training and inference.

2.3   Camera-based 3D Detection

Camera-based 3D perception tasks have gained substantial attention in recent research, largely due to the convenience and cost-effectiveness of cameras as data collection sensors. Initial efforts, such as FCOS3D [26] and DETR3D [28], explore the transition from 2D to 3D predictions. BEVFormer [15] represents a significant advancement in this area, transforming images captured from vehicle-mounted cameras into a bird's eye view (BEV) representation. This technique not only enhances a vehicle's environmental understanding but also finds applications in various downstream tasks like BEVStereo [14] and BEVDet [7] and extends to 3D occupancy prediction.

3D detection tasks are important in camera-based 3D perception and also have a high similarity to 3D occupancy prediction. The primary objective of 3D detection involves estimating the position and dimensions of objects in 3D space. DETR3D [28], drawing inspiration from DETR [33], innovatively combines 3D object queries with image features, incorporating camera intrinsic and extrinsic parameters. PETR [17] and its successor, PETRv2 [18], further refine this approach by addressing the accuracy of reference point sampling and the inclusion of global information, enhancing 3D detection through historical data integration. Compared to object detection, 3D occupancy prediction offers a finer granularity, which is crucial for navigating irregular obstacles or overhanging objects. It is imperative to explore how insights from BEV representation and 3D object detection can inspire innovative solutions tailored to the specific requirements of 3D occupancy prediction.

3   Preliminary

Problem setup. We aim to provide occupancy predictions around the ego-vehicle, given only N surround-view RGB images. More specifically, we use as input current images denoted by $\mathbf{I}_i = \{I_0, I_1, ..., I_N\}$, and use as output an occupancy prediction $\mathbf{Y}_i \in \{c_0, c_1, ..., c_M\}^{H \times W \times Z}$ defined in the coordinate of ego-vehicle, where each occupancy prediction is either empty (denoted by c_0) or occupied by a certain semantic class in $\{c_1, c_m, ..., c_M\}$. We assume known camera intrinsic parameters $\{K_i\}$ and extrinsic parameters $\{[R_i|t_i]\}$ in each frame. We assume to know whether each area is visible or not by applying a camera visibility mask.

Points of Interest. In our model, we innovatively introduce the concept of Points of Interest (PoIs), which is a set of sparse points to represent the 3D scene. PoIs can flexibly represent objects or regions that require extra attention such as pedestrians or regions near ego-vehicles and can be designed as needed in both training and inference phases. We use three types of PoIs and the definitions and functions are introduced as follows:

(1) *Standard Grids:* By sampling center points of grids in the inference stage and making predictions of standard 3D occupancy grids, our model makes a fair comparison with existing methods and achieves good results.

(2) *Adaptively Sampling:* During the training stage, our model adaptively samples points and oversamples points around them to enhance accuracy. Recognizing that volume-based methods uniformly treat all locations, our point-based approach allows for resampling in either areas of special interest or those that are challenging to learn. This adaptively resampling strategy is also used to augment the performance of volume-based methods. Consequently, our method can function as a versatile plugin, seamlessly integrating with and enhancing existing volume-based approaches.

(3) *Manually Sampling:* Our model excels in its flexibility to sample any area, specifically catering to the unique demands of various downstream tasks.

Our model can make predictions of areas beyond the standard perception range by setting PoIs to areas outside the scene manually, *e.g.* 200 meters away from ego-vehicle, which is a feat unattainable by traditional volume-based methods. This extension not only broadens the scope of inference but also introduces a new dimension to scene understanding.

These PoIs are the foundation of our method, offering precision and high flexibility.
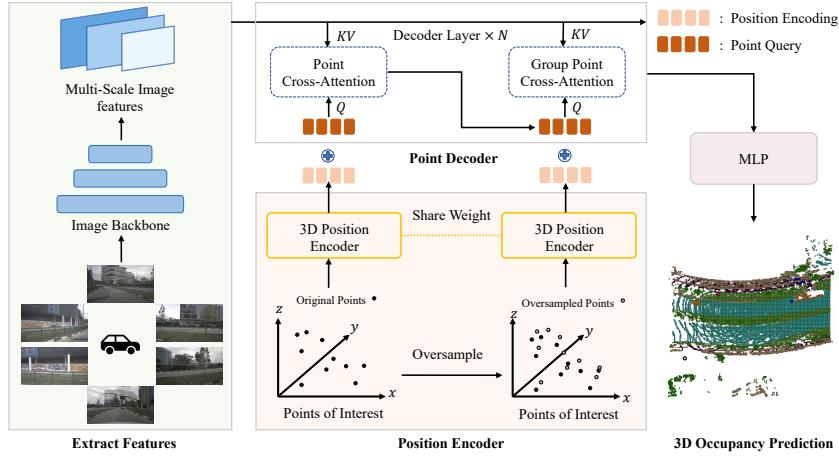
4 Method

Fig. 2: Overall framework of Occupancy as Set of Points. OSP leverages the Transformer architecture to derive 3D point features from 2D images to make 3D occupancy predictions. Initially, we extract 2D features from multi-view images. Following this, we employ a set of 3D point queries to index these 2D features. The selection of these 3D point queries depends on the Points of Interest (PoIs).

4.1 Overall Architecture

Fig. 2 shows the overall architecture of the proposed Occupancy as Set of Points. Given the images $\mathbf{I} = \{I_i, i = 1, 2, ..., N\}$ from N views, We feed these images into ResNet [6] to obtain their features.

OSP begins with sampling a set of 3D points in the space, which are the initial PoIs. In our experiments, we use center points of grids as the initial PoIs to make a fair comparison with traditional volume-based methods and provide a reliable baseline for performance comparison and evaluation. We sample $K(K = 8000)$ points within the camera's visible region and introduce random perturbations to these points.

Then we normalize the 3D points. To these normalized coordinates, we apply sine and cosine functions as a form of positional encoding. This encoded positional information is then utilized to create query position embeddings. In the training phase, within each decoder layer, the query position corresponding to each individual query remains consistent. The 3D coordinates, along with the camera's intrinsic and extrinsic parameters, are used to map these points onto the pixel plane. This mapping process yields corresponding key and value pairs. Subsequently, we employ point cross-attention mechanisms to compute the output. Then we adaptively oversample a group of M points ($M = 8000$) whose coordinates are calculated by a linear layer and employ group point cross-attention to fuse the features of the additional sampling points.

4.2   3D Position Encoder

After obtaining 3D points by applying PoIs, we first normalize coordinate points using equations as follows:

$$x, y, z = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \frac{y - y_{\min}}{y_{\max} - y_{\min}}, \frac{z - z_{\min}}{z_{\max} - z_{\min}}, \tag{1}$$

where $[x_{\min}, y_{\min}, z_{\min}, x_{\max}, y_{\max}, z_{\max}] = [-40m, -40m, -1m, 40m, 40m, 5.4m]$, are the preset boundaries of the scene. Through sine and cosine functions, we encode the normalized coordinates into high-dimensional positional information. Then we use a small MLP containing two linear layers and one ReLU layer to transform the high-dimensional positional information into learnable embeddings.

4.3   Point Decoder

We used three decoder layers to build our point decoder. Each layer of the decoder contains point cross-attention (PCA) and group point cross-attention (GPCA).

The purpose of PCA and GPCA is to integrate position embedding with image features, thereby facilitating a more cohesive representation. Owing to the high efficiency of **deformable attention** [15], our point decoder mechanisms employ this technique. Consequently, each query within our framework can be updated as follows:

$$\mathtt{DA}(\mathbf{q}, \mathbf{p}, \mathbf{F}) = \sum_{s=1}^{N_s} \mathbf{A}_s \mathbf{W}_s \mathbf{F^{2D}}(\mathbf{p} + \Delta\mathbf{p}_s), \tag{2}$$

where N_s represents the number of sampling offsets, $\mathbf{A}_s$ represents the learnable attention weights, $\mathbf{F^{2D}}(\mathbf{p} + \Delta\mathbf{p}_s)$ represents the image features collected at location $\mathbf{p} + \Delta\mathbf{p}_s$ in which $\Delta\mathbf{p}_s$ represents the offset apply to position $\mathbf{p}$.

Point cross-attention. The point cross-attention mechanism begins by taking point queries, initially set to zero, and combines them with point position encoding derived from our 3D position encoder. This combination forms the input queries. Subsequently, these queries undergo deformable cross-attention with 2D image features. Given that not all 3D points project onto the image plane, particularly in the context of the six surround views provided by the nuScenes Dataset. Each point is likely to map onto only one or two of these views. We utilize the camera's intrinsic and extrinsic parameters to ascertain which images a given point can map to. This approach ensures that for any specific point, we only consider the features of the image(s) it maps onto, significantly reducing memory consumption. As we directly derive 3D points and generate point query $\mathbf{q}$ through our 3D Position Encoder, the projection mapping of 3D points onto 2D image features is efficiently executed. $\mathbf{F}^{2D} = \{\mathbf{F}_t^{2D}, \mathbf{F}_{t-1}^{2D}, ...\}$ represents the mapped 2D image features where t indexes the images. Therefore, the formula of the point cross-attention can be described as follows:

$$\texttt{PCA}(\mathbf{q}, \mathbf{F}^{2D}) = \frac{1}{|\mathcal{V}_t|} \sum_{t \in \mathcal{V}_t} \texttt{DA}(\mathbf{q}, \mathcal{P}(\mathbf{p}, t), \mathbf{F}_t^{2D}), \tag{3}$$

where $\mathcal{V}_t$ represents the hit image, t indexes the images, $\mathcal{P}(\mathbf{p}, t)$ represents the projection function of input position $\mathbf{p}$.

Group point cross-attention. The group point cross-attention mechanism is aimed at solving the lack of local context in PCA since each point independently interacts with the image features. We adaptively oversample a group of M ($M = 8000$) points around our PoIs whose coordinates are calculated by a linear layer. We use the attention obtained from PCA and the 2D image features mapped by our group of points. Therefore, the formula of the group point cross-attention can be described as follows:

$$\texttt{GPCA}(\mathbf{q_g}, \mathbf{F}^{2D}) = \frac{1}{|\mathcal{V}_t|} \sum_{t \in \mathcal{V}_t} \texttt{DA}(\mathbf{q_g}, \mathcal{P}(\mathbf{p_g}, t), \mathbf{F}_t^{2D}), \tag{4}$$

where $\mathbf{q_g}$ represents attention calculated by PCA and $\mathbf{p_g}$ represents the input position of the oversampled group of points.

4.4   Loss Function

We apply class-wise cross-entropy loss and dice loss to this task. The ground truth $\hat{\mathbf{G}}_t \in \{c_0, c_1, ..., c_M\}$ represent semantic information of a group of spatial points. The class-wise cross-entropy loss can be computed by:

$$\mathcal{L}_{ce} = -\sum_{n=1}^{N} \sum_{c=c_0}^{c_M} w_c \hat{g}_{n,c} log(\frac{e^{g_{n,c}}}{\sum_c e^{g_{n,c}}}), \tag{5}$$

where n is the index of points, N is the number of selected points, c indexes class, $g_{n,c}$ is the predicted logits for the n-th point belonging to class c, w_c is a

weight for each class according to the inverse of the class frequency as in [21]. The dice loss can be computed by:

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2},$$

(6)

where p_i is the predicted probability for the point and g_i is the ground truth binary label for the point. Our final loss is the sum of the two losses:

$$\mathcal{L}_{all} = \mathcal{L}_{dice} + \mathcal{L}_{ce}.$$

(7)

5 Experiment

5.1 Experimental Setup

Dataset. We perform our experiments on the nuScenes Dataset with annotation provided by Occ3D. Occ3D-nuScenes benchmark is interested in a volume of 80.0m to the front and back of the car, 80.0m to the left and right side, and 6.4m in height. Each sample is divided as a group of 3D voxel grids with a dimension of $[200, 200, 16]$ since each voxel has a size of $[0.4m, 0.4m, 0.4m]$. There are 18 categories of occupancy semantics, of which the 18th category is empty and does not participate in the evaluation. Occ3D-nuScenes provides ground-truth semantical voxel grids by aggregate points of the static scenes and moving objects, respectively. Furthermore, Occ3D-nuScenes utilizes ray-casting-based methods to estimate camera visibility and provides camera visibility masks.

Evaluation metric. For this task, we use mIoU as the evaluation metric, which can be formulated as follows:

$$mIoU = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c}.$$

(8)

Following the definition of 3D occupancy prediction, we only evaluate our results in visible regions.

Implementation details. We input six images of the original size into ResNet101 to obtain image features. Our image backbone is pre-trained on the FCOS3D [26], which is the same backbone used in the BEVFormer baseline. Then the features will be taken by FPN to produce multi-scale feature maps. The dimension of features is $d = 256$. We stacked three decoder layers, with point cross-attention and group point cross-attention in each decoder layer. All the cross-attentions are deformable, and we sample 8 points for each reference point on the pixel plane. There is a small MLP as our point-based occupancy predictor that projects the 256 feature dimension to the number of classes. We train our model on 8 NVIDIA 3090 GPUs with 24 epochs. We use the AdamW optimizer with a learning rate of 2×10^{-4} and a weight decay of 0.01. The learning rate of the backbone is 10

Table 1: 3D Occupancy prediction performance on the Occ3D-nuScenes dataset. * means the performance is achieved by our implementation using the camera mask during training. † means the performance is achieved with a frozen BEVFormer baseline backbone.

Method	others	barrier	bicycle	bus	car	Cons. Veh	motorcycle	pedestrian	traffic cone	trailer	truck	Dri. Sur	other flat	sidewalk	terrain	manmade	vegetation	mIoU
MonoScene [4]	1.75	7.23	4.26	4.93	9.38	5.67	3.98	3.01	5.90	4.45	7.17	14.91	6.32	7.92	7.43	1.01	7.65	6.06
TPVFormer [9]	7.22	38.90	13.67	40.78	45.90	17.23	19.99	18.85	14.30	26.69	34.17	55.65	35.47	37.55	30.70	19.40	16.78	27.83
BEVDet [7]	4.39	30.31	0.23	32.26	34.47	12.97	10.34	10.36	6.26	8.93	23.65	52.27	24.61	26.06	22.31	15.04	15.10	19.38
OccFormer [32]	5.94	30.29	12.32	34.40	39.17	14.44	16.45	17.22	9.27	13.90	26.36	50.99	30.96	34.66	22.73	6.76	6.97	21.93
BEVFormer [15]	5.85	37.83	17.87	40.44	42.43	7.36	23.88	21.81	20.98	22.38	30.70	55.35	28.36	36.0	28.06	20.04	17.69	26.88
CTF-Occ [25]	8.09	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.0	28.53
SurroundOcc [29]	8.7	39.2	19.7	41.4	46.2	18.7	20.6	26.4	23.3	27.0	32.5	78.0	38.3	46.6	49.6	36.7	31.6	34.4
OpenOccupancy [27]	10.4	45.7	23.6	42.4	49.3	14.8	24.6	27.7	27.8	27.6	33.3	79.2	39.8	47.1	50.5	37.7	31.8	36.1
BEVDet-depth [8]	6.6	41.2	7.0	42.7	48.4	18.4	12.9	22.0	18.2	28.2	33.2	80.1	39.7	49.1	52.1	39.9	33.8	33.7
BEVDet-stereo [8]	8.6	45.9	14.3	46.0	51.2	**23.8**	18.9	24.1	22.3	33.6	37.9	81.5	40.5	**52.6**	**55.9**	**46.9**	**41.2**	38.0
BEVFormer* [15]	8.54	46.24	20.28	47.46	53.04	19.33	25.39	26.16	25.1	33.45	37.75	81.17	38.64	49.32	52.54	42.73	36.13	37.84
OSP	**11.2**	47.25	27.06	47.57	53.66	23.21	29.37	29.68	28.41	32.39	39.94	79.35	41.36	50.31	53.23	40.52	35.39	39.41
OSP†	8.87	46.33	21.32	47.51	53.14	19.6	26.12	26.84	26.68	33.67	37.94	81.21	39.13	49.48	52.76	42.73	36.12	38.20
BEVFormer w/ OSP†	10.95	**49.0**	**27.68**	**50.24**	**55.99**	22.96	**31.02**	**30.91**	**30.25**	**35.6**	**41.23**	**82.09**	**42.59**	51.9	55.1	44.82	38.17	**41.21**

times smaller. Besides, we conducted a separate experiment in which we trained our decoder with a frozen BEVFormer baseline backbone using the same setting above and this decoder will be used as a plugin module for the BEVFormer baseline.

Baseline methods. We compare Occupancy as Set of Points with existing methods replicated by Occ3D on Occ3D-nuScenes benchmark, including MonoScene [4], TPVFormer [9], BEVDet [7], OccFormer [32] and BEVFormer [15]. To make a fair comparison with the baseline, we provide results using BEVFormer by our implementation using the camera visible mask during the training.

5.2   Performance

Evaluation strategy. Our evaluation strategy for the method comprehensively utilized the three types of Points of Interest (PoIs) previously outlined. This approach entailed:

(1) Standard Grids. We conducted sampling at the centers of grids, enabling us to benchmark our method's mIoU against traditional volume-based methods. This comparison provided a direct assessment of our method's performance in a standard scenario.

(2) Adaptively Sampling. We adaptively oversample in our training phase to enhance our performance. Furthermore, we integrated our method as a complementary tool with volume-based approaches. By adaptively selecting PoIs that need refinement, our method enhances and augments existing volume-based techniques as shown in Fig. 3.

(3) Manually Sampling. We manually select points out of the standard perception range and test out the model's perceptual ability.

Through these varied and thorough evaluation techniques, we are able to comprehensively assess the performance and flexibility of our method across different scenarios and use cases

Standard grids. In Tab. 1, we benchmarked our method against existing camera-based 3D occupancy prediction techniques on the Occ3D-nuScenes benchmark by setting PoIs to standard grids. Our point-based approach achieved a notable improvement, **obtains 1.57 mIoU performance gain** compared to our implementation of the Bevformer baseline. Notably, our method outperformed the BEVFormer baseline across almost all category metrics, demonstrating a particularly clear advantage in detecting small targets, such as the bicycle ($20.28 \rightarrow 27.06$), motorcycle ($25.39 \rightarrow 29.37$), pedestrian ($26.16 \rightarrow 29.37$) and traffic cone ($25.1 \rightarrow 28.31$). This advantage is mainly because the direct sampling of spatial points is beneficial for feature extraction and mapping of small objects.

Adaptively Sampling. (1) The results shown in Tab. 4 demonstrate that our adaptively oversampling strategy improved the results from 38.01 to 38.48. (2) In the adaptive refinement experiment, we use our own implementation of the BEVFormer baseline as a representative of volume-based methods and enhance it with our method using different scales of PoIs by adaptively selecting locations with low confidence. Our model is trained with the frozen BEVFormer baseline backbone. The experimental results, as shown in Tab. 3, confirm that our plugin model significantly enhances BEVformer's performance. Notably, the BEVFormer w/ OSP results surpass those obtained from the two models operating independently. Furthermore, as we increase the size of the refined scene, the mIoU scores of BEVFormer w/ OSP also exhibit marked improvements, indicating that our refinement approach is very effective across various areas.

Manually Sampling. The Occ3D-nuScenes dataset offers annotations across an $[80m \times 80m]$ range. We manually set the points and annotations within a smaller $[60m \times 60m]$ area and set PoIs to points out of it. To comprehensively evaluate our method, we conducted assessments across three distinct ranges: the standard $[60m \times 60m]$ range, the full dataset range of $[80m \times 80m]$, and an intermediate-range spanning from $60m$ to $80m$. Results shown in Tab. 2 demonstrate the capability of our method to make predictions at manually selected locations which is outside the predefined range in this case.
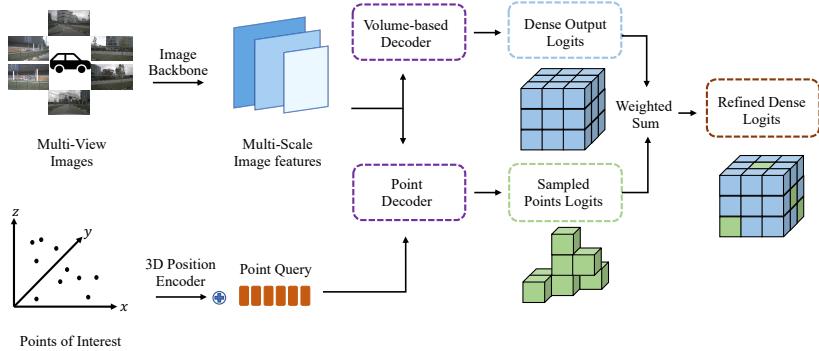
Fig. 3: Pipeline of refining volume-based methods with OSP. Given RGB images, 2D features are extracted by the frozen image backbone of the volume-based method which in our case is the BEVFormer baseline. We use the volume-based decoder to infer the entire scene, the point decoder to infer our selected 3D points, and combine the results of both using a weighted sum method

Table 2: Predict regions beyond the scene. The model was trained on scenes within a $60m \times 60m$ range, and evaluations were conducted on the training scenes, the entire size of the scene, and the distance from 60m to 80m.

Method	Scene	Range	mIoU
OSP	Standard	$0m \sim 60m$	42.29
OSP	Large	$0m \sim 80m$	29.00
OSP	Beyond	$60m \sim 80m$	23.00

Table 3: Refine BEVFormer with our decoder. † means our model is trained with a frozen BEVFormer baseline backbone. Refine scale refers to an area defined by the length multiplied by the width, centered around the self-vehicle. The entire scene spans a scale of $80m \times 80m$. As the refine scale increases, the mIoU also gradually increases.

Method	Refine Scale	mIoU
BEVFormer	-	37.84
OSP†	-	38.20
BEVFormer w/ OSP	$20m \times 20m$	38.20
BEVFormer w/ OSP	$40m \times 40m$	39.35
BEVFormer w/ OSP	$50m \times 50m$	40.32
BEVFormer w/ OSP	$80m \times 80m$	**41.21**

5.3   Ablation studies

Model architecture. We conducted ablation experiments on the model structure, particularly exploring variations in the number of transformer layers and the adaptively oversampling strategy. The results of these experiments are presented in Tab. 4. Even without an oversampling strategy and with just a 3-layer transformer, our method's baseline already surpasses the BEVFormer baseline, advancing from 37.84 to 38.01 in mIoU. The oversampling strategy further amplifies this improvement by enhancing the connections between spatial points, culminating in an increased mIoU of 38.48.

Table 4: Model architecture. To make a fair comparison with the baseline, we set the number of layers to three. The oversampling strategy brings 0.47 mIoU.

Layer Num	Oversample	mIoU
1	-	36.76
3	-	38.01
3	✓	38.48

2D image features. The performance of our method is heavily reliant on the quality of image features. To elucidate this dependency, we conducted ablation studies focusing on the output from the model's 'neck' component. These experiments demonstrated the substantial impact of using a multi-scale output from the Feature Pyramid Network (FPN). This architectural choice notably boosts the model's performance, as evidenced by the increase in metric scores from 38.67 to 39.41 as shown in Tab. 5.

Table 5: 2D image features. Increasing the multi-scale feature maps of images from 2 to 4 brings 0.74 mIoU.

Multi-scale Images Features	mIoU
2	38.67
4	39.41

Grid center sampling method. Our ablation experiments also investigated the technique for the grid center sampling method during training. While initially using grid center points as coordinates, we introduced a variation by adding random perturbations. The results, as detailed in Tab. 6, indicate the introduction of this disturbance to the grid center points elevates the mIoU to 38.67.

Table 6: Point Perturbation. We randomly applied perturbations of $0.1m$ to the coordinates of each point in three directions. This strategy brings 0.19 mIoU.

Point Perturbation	mIoU
-	38.48
✓	38.67

Loss design. The task of 3D occupancy prediction bears similarities to segmentation tasks. In the context of the Occ3D-nuScenes dataset, there is a notable issue of class imbalance. To address this challenge, we opted to incorporate dice loss into our optimization strategy. The effectiveness of this choice is evident in the results presented in Tab. 7, which demonstrate that the integration of dice loss benefits our method.

Table 7: Loss design. Incorporating dice loss for optimization brings 0.53 mIoU.

Dice Loss	mIoU
-	38.88
✓	39.41

Adaptive point sampling. During the inference process, we apply adaptive sampling (adaptively select points with high uncertainty) in BEVFormer w/ OSP and OSP. Adaptive sampling can reduce computational burden while maintaining good performance as shown in Tab. 8. BEVFormer w/ OSP using adaptive sampling means we refine BEVFormer by adaptively selecting points with high uncertainty (around 20%). OSP using adaptive sampling means we only forward points with low confidence to the next decoder layer, while directly outputting the results for points with high confidence in the decoder. Points of high uncertainty are defined as those with a confidence score less than a threshold of 0.9 after softmax.

Table 8: Adaptability of OSP. 'rel' denotes relative.

Method	Adaptive Sampling	Rel. Computation	mIoU
BEVFormer	-	-	37.84
BEVFormer w/ OSP	-	1.0×	39.35
BEVFormer w/ OSP	✓	0.8×	39.22
OSP	-	1.0×	39.41
OSP	✓	0.93×	39.08

Multi-View Images   Prediction   Groundtruth
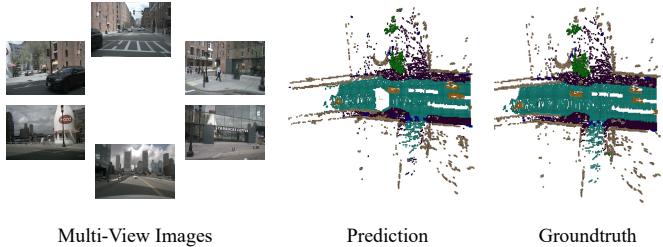
Fig. 4: Visualization of our results. Our visualization results have voids compared to the ground truth ego-vehicle position, which is due to the surrounding view having gaps near the ego-vehicle.

6   Conclusion

In this work, we present a novel perspective on 3D scene representation, viewing it through a set of points. We introduce the innovative concept of *Points of Interest* (PoIs), which significantly advances the flexibility in scene representation. Building upon the foundation laid by PoIs, we develop a highly adaptable, point-based 3D occupancy prediction framework, named OSP. We validate the strong performance and flexibility of OSP on the Occ3D-nuScenes benchmark. Our work not only contributes to the field of 3D occupancy prediction but also paves the way for more dynamic and adaptable methods in 3D scene analysis.

References

1. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9297–9307 (2019) 4
2. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020) 3
3. Cai, Y., Chen, X., Zhang, C., Lin, K.Y., Wang, X., Li, H.: Semantic scene completion via integrating instances and scene in-the-loop. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 324–333 (2021) 4
4. Cao, A.Q., de Charette, R.: Monoscene: Monocular 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3991–4001 (2022) 3, 4, 10
5. Chen, X., Lin, K.Y., Qian, C., Zeng, G., Li, H.: 3d sketch-aware semantic scene completion via semi-supervised structure prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4193–4202 (2020) 4
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 6

7. 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) 2, 3, 4, 10

8. Huang, J., Huang, G., Zhu, Z., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. CoRR (2021) 10

9. 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) 3, 4, 10

10. Jiang, H., Cheng, T., Gao, N., Zhang, H., Liu, W., Wang, X.: Symphonize 3d semantic scene completion with contextual instance queries. CoRR (2023) 4

11. Li, J., Han, K., Wang, P., Liu, Y., Yuan, X.: Anisotropic convolutional networks for 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3351–3359 (2020) 4

12. Li, J., Liu, Y., Gong, D., Shi, Q., Yuan, X., Zhao, C., Reid, I.D.: RGBD based dimensional decomposition residual network for 3d semantic scene completion. In: CVPR. pp. 7693–7702 (2019) 4

13. Li, J., Liu, Y., Yuan, X., Zhao, C., Siegwart, R., Reid, I., Cadena, C.: Depth based semantic scene completion with position importance aware loss. IEEE Robotics and Automation Letters **5**(1), 219–226 (2019) 4

14. Li, Y., Bao, H., Ge, Z., Yang, J., Sun, J., Li, Z.: Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In: AAAI. pp. 1486–1494 (2023) 4

15. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. arXiv preprint arXiv:2203.17270 (2022) 2, 3, 4, 7, 10

16. Liu, S., Hu, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., Li, X.: See and think: Disentangling semantic scene completion. In: Advances in Neural Information Processing Systems. vol. 31 (2018) 4

17. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. arXiv preprint arXiv:2203.05625 (2022) 5

18. Liu, Y., Yan, J., Jia, F., Li, S., Gao, Q., Wang, T., Zhang, X., Sun, J.: Petrv2: A unified framework for 3d perception from multi-camera images. arXiv preprint arXiv:2206.01256 (2022) 5

19. Miao, R., Liu, W., Chen, M., Gong, Z., Xu, W., Hu, C., Zhou, S.: Occdepth: A depth-aware method for 3d semantic scene completion (2023) 4

20. Moravec, H., Elfes, A.: High resolution maps from wide angle sonar. In: Proceedings. 1985 IEEE international conference on robotics and automation. vol. 2, pp. 116–121. IEEE (1985) 3

21. Roldao, L., de Charette, R., Verroust-Blondet, A.: Lmscnet: Lightweight multiscale 3d semantic completion. In: 2020 International Conference on 3D Vision (3DV). pp. 111–119. IEEE (2020) 9

22. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1746–1754 (2017) 1, 3

23. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020) 3

24. Thrun, S.: Probabilistic robotics. Communications of the ACM **45**(3), 52–57 (2002) 3

25. Tian, X., Jiang, T., Yun, L., Mao, Y., Yang, H., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving (2023) 1, 3, 10
26. Wang, T., Zhu, X., Pang, J., Lin, D.: Fcos3d: Fully convolutional one-stage monocular 3d object detection. arXiv preprint arXiv:2104.10956 (2021) 4, 9
27. 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. arXiv preprint arXiv:2303.03991 (2023) 3, 10
28. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: Conference on Robot Learning. pp. 180–191. PMLR (2022) 4, 5
29. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. arXiv preprint arXiv:2303.09551 (2023) 10
30. Zhang, J., Zhao, H., Yao, A., Chen, Y., Zhang, L., Liao, H.: Efficient semantic scene completion network with spatial group convolution. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 733–749 (2018) 4
31. Zhang, P., Liu, W., Lei, Y., Lu, H., Yang, X.: Cascaded context pyramid for full-resolution 3d semantic scene completion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7801–7810 (2019) 4
32. Zhang, Y., Zhu, Z., Du, D.: Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. arXiv preprint arXiv:2304.05316 (2023) 3, 10
33. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2020) 5
34. Zuo, S., Zheng, W., Huang, Y., Zhou, J., Lu, J.: Pointocc: Cylindrical tri-perspective view for point-based 3d semantic occupancy prediction (2023) 4