AdaOcc: Adaptive-Resolution Occupancy Prediction

Chao Chen[1]   Ruoyu Wang[1,2]   Yuliang Guo[2]   Cheng Zhao[2]   Xinyu Huang[2]   Chen Feng[1]   Liu Ren[2*]

[1]New York University   [2] BOSCH Research North America

https://github.com/ai4ce/Bosch-NYU-OccupancyNet/

Abstract

Autonomous driving in complex urban scenarios requires 3D perception to be both comprehensive and precise. Traditional 3D perception methods focus on object detection, resulting in sparse representations that lack environmental detail. Recent approaches estimate 3D occupancy around vehicles for a more comprehensive scene representation. However, dense 3D occupancy prediction increases computational demands, challenging the balance between efficiency and resolution. High-resolution occupancy grids offer accuracy but demand substantial computational resources, while low-resolution grids are efficient but lack detail. To address this dilemma, we introduce AdaOcc, a novel adaptive-resolution, multi-modal prediction approach. Our method integrates object-centric 3D reconstruction and holistic occupancy prediction within a single framework, performing highly detailed and precise 3D reconstruction only in regions of interest (ROIs). These high-detailed 3D surfaces are represented in point clouds, thus their precision is not constrained by the predefined grid resolution of the occupancy map. We conducted comprehensive experiments on the nuScenes dataset, demonstrating significant improvements over existing methods. In close-range scenarios, we surpass previous baselines by over 13% in IOU, and over 40% in Hausdorff distance. In summary, AdaOcc offers a more versatile and effective framework for delivering accurate 3D semantic occupancy prediction across diverse driving scenarios.

1. Introduction

Accurate representation of surroundings is vital for autonomous driving decision-making. The required perception granularity varies based on tasks: highways may need sparse but long-range views, while urban areas require dense, close-range detail. Finding a representation enabling safe navigation in all scenarios, adapting to dynamic changes in real-time, remains a significant challenge.

Various scene representations have emerged in au-

tonomous driving research. The prototypical ones include uniform voxel-based representations [1,18,40,44,49], bounding box representations [9,13,37,39,41,46], implicit representations [32,34,51], point-based representations [14, 33, 45], and other forms of representations [25, 28]. While object-centric representations using bounding boxes were traditionally popular, voxel-based representations have become more prevalent in recent years due to the rich information they offer in 3D semantic occupancy maps. Since each voxel contains both occupancy and semantic information, voxel-based representations provide a more comprehensive understanding of the scene. They include additional background descriptions and capture surface shapes with a certain level of granularity. Moreover, voxel-based representations are popular for their seamless integration with navigation and planning frameworks.

Despite the flexibility of 3D semantic occupancy maps in adjusting grid sizes, most of the existing methods produce relatively low resolution ($0.4\,\mathrm{m}$ or $0.5\,\mathrm{m}$ sized) occupancy grid [36,37,44], limiting their applications to highway driving. In urban driving or parking scenarios, higher-resolution representation is essential for precise vehicle maneuvering. As seen in Figure 2, the distance measured between the two cars can deviate as much as 0.6m when using 0.8m and 0.2m voxels. However, increasing resolution leads to a cubic increase in computational complexity and GPU memory consumption.

To balance memory efficiency and perception accuracy, we propose two strategies:

Non-uniform resolution. In vehicle path planning, close-range elements are considered more important than far-range elements, and objects (e.g., cars, pedestrians) are prioritized over background elements (e.g., roads, sidewalks). Therefore, we propose a non-uniform resolution representation, focusing high-resolution prediction on close-range objects.

Multi-modal 3D representation. This design is based on the two major drawbacks of voxel grid representation. Firstly, its accuracy is limited by voxel size, leading to GPU resource overload when aiming for finer granularity. Secondly, real-world scene sparsity often results in many

---

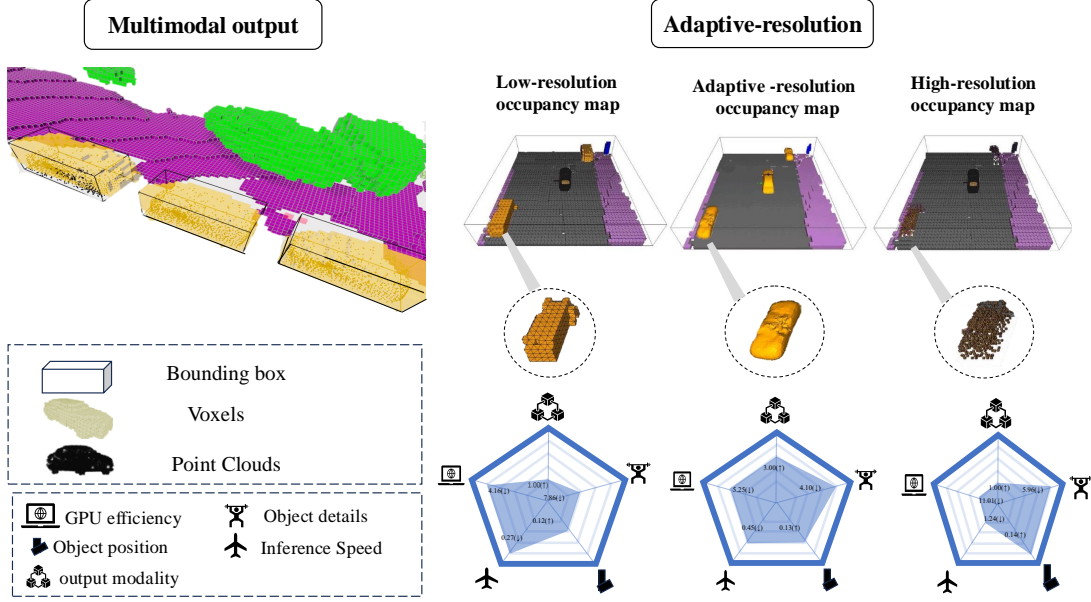*The corresponding author: Liu Ren liu.ren@us.bosch.com

Figure 1. **AdaOcc** is a multimodal, adaptive-resolution approach designed for high precision in regions of interest while maintaining overall resource efficiency. On the left, AdaOcc's outputs are visualized using bounding boxes, voxels, and point clouds. On the right, visual comparisons are made between low, adaptive, and high-resolution occupancy map predictions, evaluated across five indicators that assess both accuracy and efficiency. The adaptive-resolution approach represents the data in surface format by applying surface reconstruction to point clouds. It yields the most balanced scores, effectively accommodating diverse driving tasks while managing computational costs.
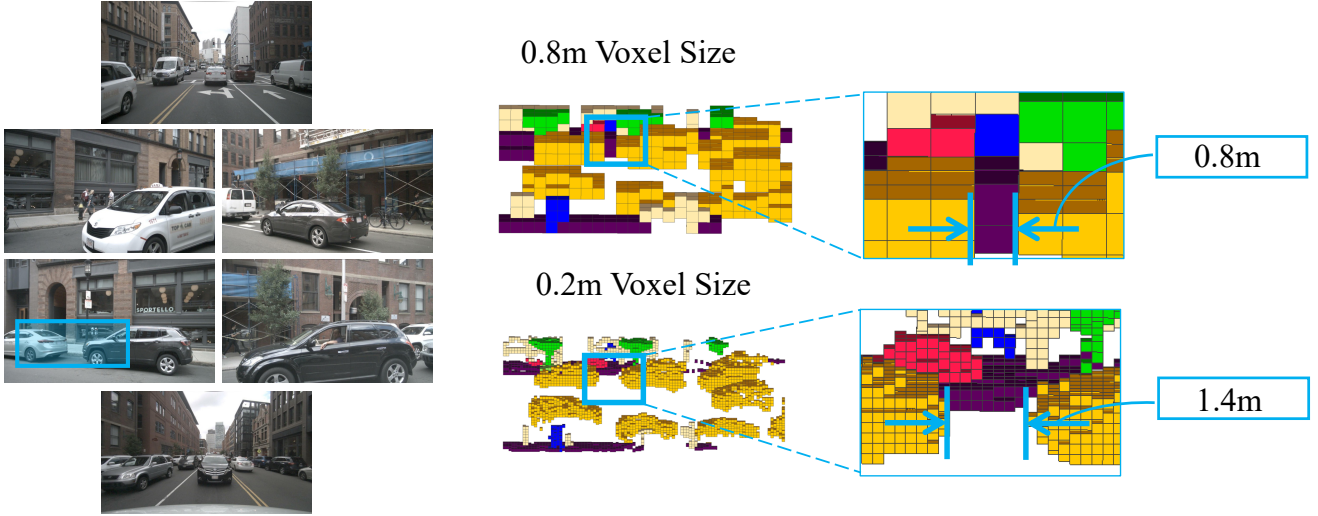

Figure 2. **Impact of Grid Resolution.** The surrounding scene captured by the RGB images can be represented in occupancy maps with different voxel sizes. The measured distance between the same two cars with grid sizes of 0.8m and 0.2m can differ by up to 0.6m (1.4m - 0.8m). This discrepancy can significantly affect precise navigation in safety-critical scenarios.

unoccupied voxels, causing inefficient memory usage. To address these issues, we propose a multimodal approach for 3D representation, incorporating outputs such as voxel grids, point clouds, and bounding boxes, rather than relying solely on voxel grids for 3D semantic occupancy (as shown on the left of Fig. 1). Among these, point clouds are

particularly advantageous due to their high detail and independence from voxel size. They do not necessarily require increased memory usage and only indicate the presence of occupied elements, thus overcoming the limitations of voxel grids.

By employing the aforementioned strategies, we intro-

duce AdaOcc, a multi-modal, adaptive-resolution semantic occupancy prediction approach. AdaOcc prioritizes precision in critical areas vital for autonomous driving decisions, i.e., close-range objects. Rather than uniformly allocating computational resources, AdaOcc employs fine-detailed reconstruction via point cloud for close-range objects while using coarser 3D voxel grid predictions for 3D semantic occupancy as supplement. We draw inspiration from object-centric methods [19, 41, 42] that incorporate an Object Proposal Network (OPN), to identify the regions of interest (ROIs) in 3D space. Subsequently, we connect a 3D point cloud decoder [47] to generate detailed point clouds within these ROIs. The right panel of Fig. 1 demonstrates that AdaOcc's adaptive-resolution representation strikes a balance of high accuracy and efficiency for autonomous driving tasks.

To enhance the synergy between different modalities, we jointly train a shared backbone (2D bird's-eye view(BEV) [19] or 3D feature volume [40]), providing voxel grids, bounding boxes and point clouds in a unified network architecture. Leveraging multi-modal representation, our occupancy prediction model can be effectively trained using only coarse ground-truth occupancy data plus raw Lidar points while still being evaluated as high resolution occupancy. AdaOcc is experimentally compared to previous methods on nuScenes dataset in both close-range and long-range scenarios to demonstrate its effectiveness. Notably, the remarkable improvement in IOU (¿13 %) and Hausdorff distance (¿ 40%) metric for close-range evaluations demonstrates our method's superior performance in capturing the details of object reconstruction and achieving precise object position estimation.

In summary, our contributions are listed as follows:

1. We propose a multi-modal adaptive-resolution method, offering three output representations with high precision in critical regions while maintaining efficiency for real-time applications.

2. We develop an effective joint training paradigm that boosts the synergy between the occupancy prediction and object folding branches.

3. Our approach demonstrates superior accuracy on the nuScenes dataset, particularly excelling in close-range scenarios that require precise maneuvering.

2. Related Works

3D semantic occupancy prediction. 3D semantic occupancy prediction is rapidly evolving, playing a crucial role in achieving precise perception necessary for safely navigating urban environments. Several pioneering works [1, 18] are designed for single-view image inputs, laying the foundation for dense geometry and semantic inference from a singular perspective. In contrast, other methods [10, 26, 44, 49] utilize surround-view images to achieve a comprehensive 360-degree understanding of the environment. Among these, OpenOccupancy [40] provides a benchmark to evaluate occupancy prediction at the finest level, using a 0.2-meter voxel size. The proposed method, CONet, was the first to practically realize occupancy prediction at a 0.2-meter voxel scale through a cascaded approach.

Since these approaches rely on uniformly sampled voxels, their precision is largely constrained by the total number of voxels a computing unit can afford. By focusing computational resources on target objects, AdaOcc achieves highly precise perception in critical regions while maintaining the overall computation cost.

3D object detection from surround-view images. The landscape of camera-based surround-view 3D object detection in autonomous driving has seen significant advancements in unified framework design, as demonstrated by [9, 22, 46, 48]. Researchers have concentrated on transforming multiple perspective views into a unified 3D space within a single frame, as demonstrated by studies such as [9, 13, 19, 22, 42, 43]. This process can be categorized into two main approaches: (1)BEV-based methods [8, 9, 13, 16, 17, 19, 37, 41, 46], (2) sparse query based methods [2, 20, 22, 42, 43]. In comparison, BEV methods are considered more compatible with other 3D perception tasks demanding dense outputs, such as 3D occupancy prediction, depth estimation, and 3D scene reconstruction.

Inspired by [37, 41], AdaOcc further enhances 3D comprehension by integrating object detection, occupancy prediction, and object surface reconstruction into a unified framework. Our framework not only comprehensively represents the entire scene but also focuses on high surface precision within object regions. While [41] employs a similar strategy by performing semantic occupancy prediction within certain ROIs, it still outputs 3D voxels in uniform grids. This approach continues to face the efficiency-precision dilemma in choosing the grid resolution, as seen in other occupancy prediction methods.

2D-3D encoding backbones. Within the realm of 2D-3D encoding backbones, two predominant methodologies emerge: transformer-based backbones [19, 23, 24, 29] and Lift-Splat-Shoot (LSS)-based backbones [18, 30, 40]. Transformer-based backbones typically create a query grid in 3D space, project these grid points onto 2D image planes, and then aggregate the extracted features back to the query grid using a deformable transformer [50]. Conversely, LSS-based backbones incorporate a depth probability prediction module that allocates 2D image features across 3D space according to estimated depth probabilities. Each approach offers distinct advantages. For our experiments, we chose BEVFormer [19] (transformer-based) and CONet

[40] (LSS-based) as the baseline networks to underscore the capabilities of AdaOcc.

Multi-resolution 3D representations. Multi-resolution representations have gained considerable traction across various fields in computer graphics and geometric modeling, as evidenced by seminal works [5,6,15,27]. Several approaches [4,7,12,38] adopt a hierarchical method for shape reconstruction, starting with a preliminary low-resolution model that is progressively refined into a high-resolution output. Other methods [21, 35] extend hierarchical structures, such as octrees of implicit functions, to represent the radiance field for neural rendering, yet the granularity of the octree is predetermined by the depth map input In contrast, both MDIF [3] and FoldingNet [47] offer representations of object shapes with adjustable levels of detail.

Within the field of occupancy prediction, CONet [40] pioneers a coarse-to-fine strategy, refining only the occupied areas in the coarse occupancy map to achieve the first practical 0.2-meter semantic occupancy prediction method. Building solely upon the coarse occupancy map of CONet, our approach significantly improves performance in terms of Hausdorff distance and reduces memory usage, as detailed in Section 4.

3. Methodology

Problem statement. We formulate our task as a multi-modal, adaptive-resolution occupancy prediction. The input of the network is a set of surround-view input images $\mathcal{I} = \{I_1, I_2, ..., I_N\}$, and the output of the network are in **multiple modalities**, including: (1) 3D semantic occupancy map $M \in \mathbb{R}^{H \times W \times D}$, where this map spans from $X \subset [x_{min}...x_{max}]$, $Y \subset [y_{min}...y_{max}]$, $Z \subset [z_{min}...z_{max}]$. (2) a set of bounding boxes represented by translation(x_i, y_i, z_i), rotation(qx_i, qy_i, qz_i, qw_i), sizes (h_i, w_i, d_i), and (3) object shape in point cloud format $\mathbb{R}^{N \times K \times 3}$, here we pick K = 2500.

By means of **adaptive-resolution**, we aim to create a mixed-resolution occupancy map that combine fine resolution for objects with coarse resolution for all matters. The grid size for the occupancy map includes 0.2m, 0.4m, 0.8m, and so on. In this work, we define high resolution as which the voxel size is less than or equal to 0.2m. Otherwise, it is low resolution.

Architecture overview. Our approach is versatile, capable of integrating with either BEVFormer [19] or CONet [40], as depicted in Figure 3. It processes six surround-view input images $\mathcal{I}$ through a 2D-3D encoder. Specifically, images $\mathcal{I}_t$ captured at time t are processed using a CNN to extract 2D image features. These features are subsequently projected into a 3D feature volume F that facilitates semantic occupancy prediction, object detection, and object surface reconstruction. The BEV feature is considered a specific instance of the 3D feature volume.

3.1. Occupancy Decoder

Ego-centric occupancy perception is designed to create semantic occupancy maps of fixed grid size in surround-view driving scenarios. This module is intended to provide a holistic understanding of the entire area, allowing for the use of a low-resolution occupancy decoder for greater efficiency. The surrounding occupancy labels, denoted by V_t, are predicted through:

$$V_t = MLP\left(F_{t-1}, F_t\right), \tag{1}$$

where using BEVFormer as the framework, the previous and the current feature volume ($F_{t-1}, F_t \in \mathbb{R}^{H \times W \times D \times C_{vox}}$) are fed into an MLP-based voxel decoder, to get the coarse 3D semantic occupancy prediction ($V_t \in \mathbb{R}^{H \times W \times D}$), C_{vox} represents feature dimension. All CONet variants(using CONet as backbone) only relies on the current feature (F_t) to predict V_t and then applies an additional step of attention-based occupancy refinement on top of the original occupancy prediction.

3.2. 3D Object Detector

The 3D object detector is designed to generate 3D object bounding boxes that facilitate object-centric shape reconstruction or can be directly used for downstream tasks. The predicted 3D object bounding box, denoted as $\hat{B}$, is defined as follows:

$$\hat{B} = \arg\max_B P(\text{object}|B) \cdot P(B|I), \tag{2}$$

where $P(\text{object}|B)$ is the probability that an object is present given the bounding box B, $P(\text{object}|B)$ is the likelihood of the bounding box B given the input image I. We follow the setup from DETR to regress 900 bounding boxes and compute their object classification scores. The accuracy of the 3D object detector is critical for detailed surface reconstruction, as it directly influences the quality of the results.

3.3. FoldingNet Decoder

The FoldingNet decoder processes 3D features and predictions, producing finely detailed surfaces for targeted objects. Initially, FoldingNet [47] utilizes PointNet [31] to encode an object's point cloud and then decodes the latent features into another point cloud at an arbitrarily selected resolution. Adapting this process to effectively leverage existing 3D features within our unified framework to directly output highly accurate surface points introduces distinct challenges. Moreover, driving scenarios often involve only partial observations of an object; the variability in partial visibility combined with inaccuracies in the predicted
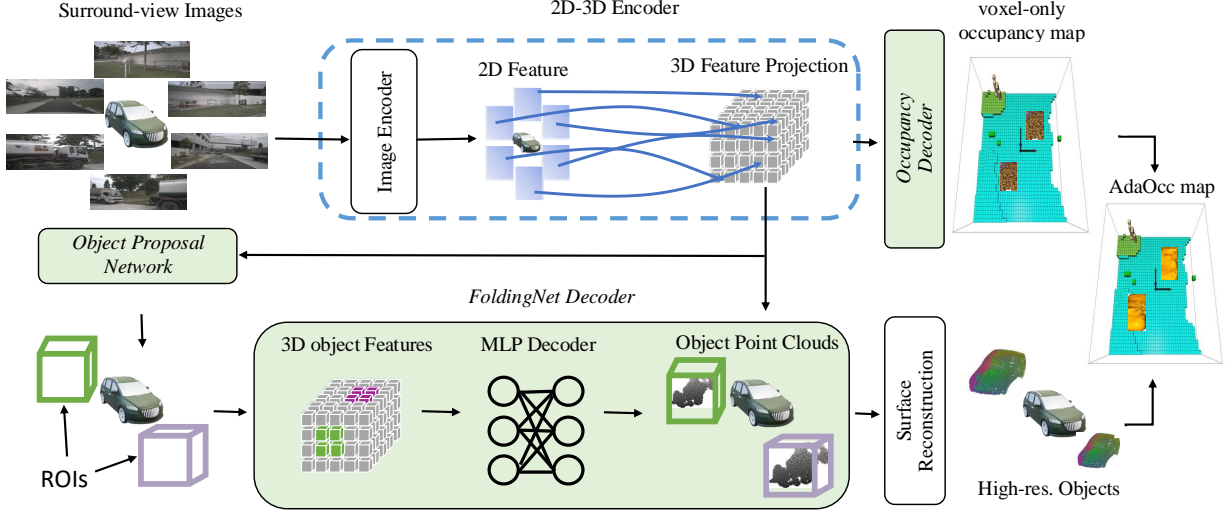
Figure 3. **AdaOcc Pipeline.** We combine a low-resolution occupancy map and a high-resolution object point cloud to create a adaptive-resolution map. The green boxes represent three outputs from our spatial-temporal encoder: occupancy prediction, 3D object detection, and point cloud reconstruction. Depending on the backbone, the 2D-3D encoder varies. BEVFormer projects image features to a BEV feature volume, while CONET projects them to a 3D feature volume. We consider a BEV feature as a special case of a 3D feature volume where depth dimension equals to 1.

object boxes can further complicate the object-centric decoding process.

Box-Aligned Object Feature Aggregation. Leveraging our object proposal network, we can concentrate on reconstructing the point cloud for each object individually. We establish a regular sampling grid $G = p_i, p_i \in \mathbb{R}^3$ within each provided 3D bounding box. The sampling grid is initially created in the object coordinate frame, in alignment with the three dimensions of the bounding box. Then G is transformed to the ego-vehicle coordinate frame via the object pose T. We retrieve the set of feature vectors from the 3D feature volume F using G and apply max pooling over all sampled features to obtain an "object feature vector" $\mathbf{c}$. If features appear at a floating-point location, cubic interpolation is employed to retrieve the feature. The insight behind the sampling process is that *the features within a bounding box encodes the local surface shape related to the object.* The max pooling operation enhances the robustness of $\mathbf{c}$ to errors in bounding box prediction and to issues of partial visibility. The sampling process can be represented using the following equation:

$$\mathbf{c} = \text{maxPooling}(\{F(T(p_i))\}), \ p_i \in G. \quad (3)$$

In training, we use ground truth bounding boxes poses to transform the 3D sampling grid, while in testing we use predicted bounding boxes poses.

Point cloud decoding. After the feature encoding the object shape is retrieved, the object surface point cloud $\mathbf{P}$ is decoded as:

$$\mathbf{P} = f_\theta(\mathbf{c}, \mathbf{g}), \quad (4)$$

where $\mathbf{g}$ is the 2D sampling grid used by the FoldingNet decoder f_θ, which is a multi-layer perceptron (MLP) to decode the point cloud.

3.4. Joint training and losses

In our approach, we address the challenge of object-centric occupancy prediction through the integration of three components. A joint training paradigm can effective enhance the synergy between different modules. The effectiveness of the joint training is further validated via ablation study in supplemental materials.

Our joint training approach integrates a combination of losses from various modules: semantic occupancy loss, object detection loss, and surface reconstruction loss. The **semantic occupancy loss** ($\mathcal{L}_{\text{sem}}$), which utilizes focal loss, is designed for predicting semantic occupancy within a fixed grid size. For object detection, we employ the **object detection loss** ($\mathcal{L}_{\text{det}}$), incorporating both focal loss for classification and L1 loss for the regression of bounding boxes. This loss function not only selects N valid boxes from a pool of candidates but also accurately estimates the position of each box simultaneously. Furthermore, the **surface reconstruction loss** ($\mathcal{L}_{\text{surf}}$), using chamfer loss, is applied to the surface reconstruction of foreground objects to ensure precise alignment between the predicted and actual object point clouds. These three loss functions collectively enhance the efficiency of our adaptive occupancy prediction framework. More detailed descriptions of these loss components can be found in the supplementary materials.

4. Experiment

We conducted extensive experiments using the NuScenes dataset, which included evaluations of occupancy predictions at both close and full ranges, as well as object detection. Because comparing voxelized ground truth with multi-modal output presents challenges, we converted our detailed object point clouds into an occupancy representation with a grid size of 0.2m for evaluation purpose. This resolution matches the ground truth voxel size in CONet and OpenOccupancy [40].

4.1. Dataset

Our experiments engage NuScenes dataset to assess our object-centric occupancy prediction methods. In our experimental configuration, the ground truth labels encompass a bounded range in the x-direction from -50.0 to +50.0 meters, in the y-direction from -50.0 to +50.0 meters, and in the z-direction from -5.0 to 3.0 meters. Furthermore, to evaluate a approach's performance under various voxel resolutions, we partition the space into voxels with granularity settings of 0.2 meters, 0.4 meters, and 0.8 meters. Given above experimental setups, we evaluate candidate methods within the varying defined spatial boundaries and at different voxel resolutions. Note that the ground truth 3D semantic occupancy comes from [40]. Note that our setup is similar to the OpenOccupancy Benchmark [40], except that we retain very small object ground truth (GT) boxes in both our training and validation sets.

4.2. Baselines

In this paper, we selected BEVFormer [19] and CONet [40] as state-of-the-art baselines for our evaluation. We aim to enhance these baselines with a multi-resolution representation to demonstrate the robustness and flexibility of our approach in improving occupancy prediction accuracy across various methods. BEVFormer, which utilizes object detection for scene representation, proves its efficacy in occupancy prediction tasks [37]. We aim to enhance BEVFormer's accuracy through the adoption of a more granular representation method, specifically FoldingNet. Conversely, CONet employs Depth Net for initial rough occupancy predictions, subsequently refined using a transformer. Although its refinement process is conceptually similar to our approach, it lacks flexibility and does not efficiently utilize GPU resources, as it refines all occupancy grids uniformly. Our method focuses on refining predictions particularly for nearby objects, optimizing the overall resource expenditure.

4.3. Evaluation metric

We assess the performance of our object detection, and occupancy prediction approach respectively. For object detection, we follow exact same procedure as OccNet [37], shown in Table. 4. For occupancy prediction, the overall occupancy is evaluated by Intersection over Union (IOU), and per-class occupancy is evaluated by mean Intersection over Union (mIOU). In addition, we apply *Hausdorff Distance* [11] for detailed assessment of the accuracy of object shapes. It evaluates the similarity between the predicted object point cloud and the ground truth object point cloud by measuring the maximum distance between them after bipartite matching. Hausdorff distance is expected to more adequately describe the precision of object shapes, not only their positional accuracy. Detailed descriptions of the evaluation metrics are provided in the supplementary materials. For 3D voxel grid, we use the centers of voxels as the points to calculate Hausdorff distance. In our experiments, We calculate the Hausdorff distance against ground truth voxel grid using the output point cloud in AdaOcc, while using the finest voxel grid they can get in the baseline methods.

4.4. Close-range occupancy prediction

This section evaluates close-range occupancy predictions, crucial for narrow path navigation and parking in autonomous driving.

Without lose of generality, we define close-range as spanning from -12.8 to +12.8 meters in both x and y directions, and from -5.0 to 3.0 meters in the z-direction. Depth estimation within 30 meters has been shown to be accurate [39], allowing for precise predictions and enhanced model performance in this range. We train the AdaOcc-model with two backbones from BEVFormer and CONet, named AdaOcc_B and AdaOcc_C, respectively. We also use three grid resolutions (0.2m, 0.4m, and 0.8m) in training and assess all models by upscaling the grid resolution to 0.2m for Intersection over Union (IOU) calculations. The focus in close-range settings is on obstacle avoidance, and therefore, mIOU is not included. Furthermore, IOU results for BEVFormer are only provided at 0.4m and 0.8m training grid sizes due to memory constraints on GPUs like the RTX 3090.

Close-range IOU on BEVFormer. Tab. 1 shows that AdaOcc based on BEVFormer consistently demonstrates an IOU improvement of at least 13% for train grid size of 0.4m and 0.8m. Additionally, since BEVFormer already includes an object detection framework, incorporating object surface reconstruction does not add significant overhead in either evaluation time or GPU usage. This ensures that using the folding method on BEVFormer provides a lightweight and efficient way to improve close-range IOU.

Close-range IOU on CONet. For CONet, AdaOcc still shows some improvement. However, AdaOcc_C demonstrates significant improvement at a coarse training grid size (0.8m), but its performance is not as good at a fine training grid size (0.2m). This is because CONet's inherent coarse-to-fine refinement mechanism already provides substantial improvement in occupancy prediction. Adding extra object
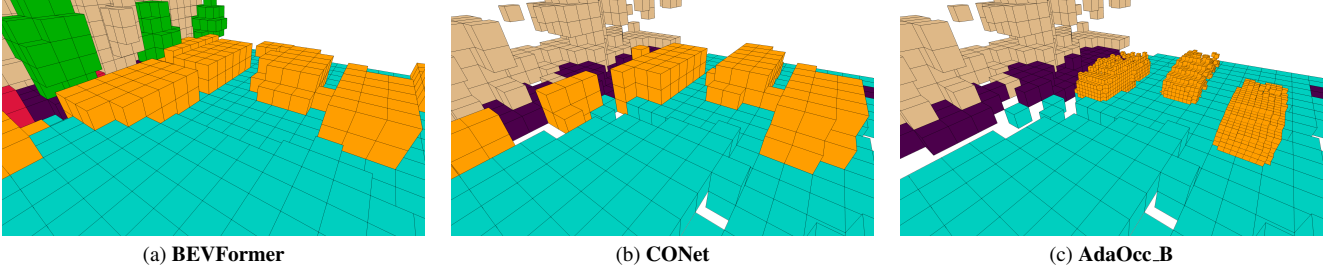
| (a) **BEVFormer** | (b) **CONet** | (c) **AdaOcc_B** |

Figure 4. **Qualitative Results on BEVFormer, CONet, and AdaOcc_B**. As observed, BEVFormer and CONet involve erroneous connections between different objects. Benefiting from adaptive-resolution, AdaOcc distinctly separates each car with clear margins.

Table 1. **Close-range occupancy evaluation.** We evaluate occupancy prediction in Hausdroff Distance, IOU, evaluation time per iteration and GPU usage on four methods: BEVFormer, CONet, AdaOcc_B(AdaOcc based on BEVFormer), and AdaOcc_C (AdaOcc based on CONet). The close range spans from -12.8 to +12.8 meters in both x and y directions, and from -5.0 to 3.0 meters in the z-direction. Colored numbers indicate the percentage of improvements over baselines. Each model is trained at different resolutions but evaluated at the same 0.2m grid size.

Method	Train Grid Size(m)	Hausdorff Distance(m)(↓)	IOU(↑)	Eval. Time(s)	GPU Usage(GB)
BEVFormer	0.4	7.868	0.125	0.443	5.186
	0.8	—	0.122	0.272	4.166
CONet	0.2	10.816	0.243	0.383	16.310
	0.4	—	0.192	0.367	9.476
	0.8	—	0.170	0.292	8.768
AdaOcc_B	0.4	**4.099(+47.9%)**	0.142(+13.6%)	0.455	5.250
	0.8		0.140(+14.75%)	0.315	4.171
AdaOcc_C	0.2	**5.967(+44.8%)**	0.246(+1.2%)	1.348	18.274
	0.4		0.197(+2.6%)	1.239	11.010
	0.8		0.193(+13.5%)	0.770	10.314

detection and object surface reconstruction at a fine resolution does not significantly enhance the original results. Additionally, CONet is a resource-intensive method as it refines every coarsely occupied cell. While object surface reconstruction brings some benefits, the extra object detection head and foldingnet head make the gains of AdaOcc at training grid sizes of 0.4m and 0.8m not particularly worthwhile.

Qualitative analysis. Figure. 4 demonstrates that the classic occupancy prediction methods are likely to connect different objects of the same class which will inevitably influence the close-range path planning performance. In contrast, the overall reconstruction quality of AdaOcc for each object is remarkably better than other baselines within a given range. The solution to poor object detection is to conduct object detection and reconstruct the object at a close range. More qualitative comparisons are included in the supplemental materials.

Analysis of Hausdorff distance. As discussed, AdaOcc based on all baselines has the best average Hausdorff distance. However, we have observed that the misinterpretation of bounding box positions and categories, especially for small objects like humans and bicycles, can significantly impact the accuracy of close-range object occupancy

Table 2. **Full-range occupancy evaluation.** In this context, *full-range* refers to the area where x ranges from -50.0m to 50.0m, y ranges from -50m to 50m, and z ranges from -5m to 3m. Besides the additional mIOU to provide a more comprehensive per-class measure of object IOU, other annotations follow Table. 1.

Method	Train Grid Size(m)	IOU(↑)	mIOU(↑)
BEVFormer	0.4	0.122	0.072
	0.8	0.089	0.053
CONet	0.2	0.156	0.095
	0.4	0.136	0.082
	0.8	0.120	0.074
AdaOcc_B	0.4	0.128(+4.9%)	0.089(+23.6%)
	0.8	0.093 (+4.4%)	0.089(+67.9%)
AdaOcc_C	0.2	0.157(+0.6%)	0.093 (-2.1%)
	0.4	0.136 (+0.0%)	0.085(+3.6%)
	0.8	0.122 (+1.6%)	0.079(+6.8%)

prediction tasks. Thus, it explains why we generate an occupancy map by integrating the voxelized object occupancy onto the coarse occupancy map, instead of replacing one by another. We believe that voxelized object occupancy is a good reference and complement to the coarse occupancy map, especially when the coarse occupancy map misinterprets the occupied cell as unoccupied.

Evaluation time on different voxel grid sizes. BEVFormer has the shortest evaluation time because it predicts

Table 3. **Full-range per-object-class occupancy evaluation.** Our AdaOcc variants are compared with prior state-of-the-art methods in terms of IOU for each object class at a single grid resolution of 0.8m.

Class Name	Baselines	Overall	Barrier	Bicycle	Bus	Car	Construction	Motorcycle	Pedestrian	Traffic Cone	Trailer	Truck
IOU	BEVFormer	0.053	0.014	0.035	0.077	0.095	0.050	0.067	0.074	0.037	0.027	0.061
	CONet	0.074	0.020	0.073	0.118	0.126	0.057	0.067	0.097	0.054	0.056	0.072
	AdaOcc_B	**0.089**	**0.029**	**0.082**	**0.144**	**0.159**	**0.087**	**0.103**	**0.117**	**0.066**	0.061	0.044
	AdaOcc_C	0.079	0.028	0.071	0.141	0.128	0.077	0.072	0.049	0.041	**0.105**	**0.080**

Table 4. **Object detection evaluation.** Since our work is mainly constructed from BEVFormer and CONet respectively, we compare to those 3D detectors involved in the two baselines or share a similar idea.

Method	mAP↑	NDS↑	mAOE↓	mAVE↓	mAAE↓	mATE↓	mASE↓
BEVFormer	0.271	0.390	0.578	0.541	0.211	0.835	0.293
VoxNet	**0.277**	0.387	0.586	0.614	0.203	0.828	**0.285**
OccNet	0.276	0.390	0.585	0.570	**0.190**	0.842	**0.285**
AdaOcc_B	0.273	**0.391**	**0.577**	0.574	0.222	**0.808**	0.295
AdaOcc_C	0.272	0.390	0.579	**0.532**	0.209	0.833	0.291

every occupancy grid equally and coarsely. In contrast, CONet requires the most time for evaluation due to its two-stage process. Initially, it computes a coarse occupancy map similar to BEVFormer, and subsequently, it refines each grid for all occupied grids. As CONet itself lacks an object detection pipeline, its cost for object surface reconstruction will be higher than BEVFormer's, as it requires both an object detection and a foldingnet head for object point cloud reconstruction.

4.5. 3D object detection

3D detection task with 3D box regression coarsely regresses the location of the foreground object can be coarsely regressed. In this section, we prove that the joint training of occupancy prediction, 3D detection, and surface reconstruction can improve the detector performance for all three models (BEVNet, VoxNet, and OccNet) [37], in terms of mAP, NDS, and other parameters. We developed the object detection pipeline from BEVNet and CONet, and the performance of the object detection method is very similar to the other three baseline methods.

4.6. Full-range occupancy prediction

The adaptive-resolution approach represents a computationally efficient and flexible strategy that strikes a balance between accuracy and efficiency. It involves generating a full-range adaptive-resolution occupancy map by incorporating close-range voxelized object occupancy onto a full-range coarse occupancy map. Similar to Sec. 4.4, we tested AdaOcc with grid resolution=0.2m, 0.4m, and 0.8m.

Full range IOU and mIOU. Tab. 2 illustrates that AdaOcc based on both baselines outperforms their baseline model at grid resolutions of 0.2m, 0.4m, and 0.8m. However, we can observe that globally, AdaOcc based on BEV-Former demonstrates more significant improvements. This is consistent with the findings in Section 4.4. CONet already has some refinement at training grid sizes of 0.2m and 0.4m, so additional object surface reconstruction does not yield significant improvements.

Per-object-class evaluation in IOU. In our occupancy prediction evaluation, we prioritize object segments over static scenes, particularly for classes like pedestrians and vehicles, as their movements are unpredictable. Similarly, Table 3 reveals that AdaOcc surpasses the other baselines for all 10 object classes.

Through examining mAP and NDS for BEVFormer and AdaOcc_B shown in Tab. 4, we establish that AdaOcc slightly outperforms BEVFormer, the foundation upon which AdaOcc is built. This finding demonstrates that the object surface reconstruction task not only enhances the accuracy of occupancy prediction but also enriches the learned features for object detection.

5. Conclusion

In conclusion, our proposed approach offers a multi-modal adaptive-resolution method, providing three output representations with highly precise surfaces in critical regions, while ensuring efficiency for real-time applications. Additionally, we develop an effective joint training paradigm to enhance synergy between the occupancy and folding networks, resulting in improved near-range occupancy prediction performance. Our methods exhibit superior accuracy on the nuScenes dataset, highlighting a focus on detailed surface reconstruction.

Limitation. We observe that the joint training method does not significantly improve the quality of object detection tasks. Further investigations into the interaction between the coarse occupancy prediction and the object surface reconstruction are needed to boost the consistency between different representations. In addition, the efficiency of the unified framework can be further optimized via more advanced parallelized designs.

References

[1] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1, 3

[2] Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Chang Huang, and Wenyu Liu. Polar parametrization for vision-based surround-view 3d detection. *arXiv preprint arXiv:2206.10965*, 2022. 3

[3] Zhang Chen, Yinda Zhang, Kyle Genova, Sean Fanello, Sofien Bouaziz, Christian Häne, Ruofei Du, Cem Keskin, Thomas Funkhouser, and Danhang Tang. Multiresolution deep implicit functions for 3d shape representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13087–13096, 2021. 4

[4] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5868–5877, 2017. 4

[5] Leila De Floriani and Paola Magillo. Multiresolution mesh representation: Models and data structures. *Tutorials on Multiresolution in Geometric Modelling: Summer School Lecture Notes*, pages 363–417, 2002. 4

[6] Igor Guskov, Wim Sweldens, and Peter Schröder. Multiresolution signal processing for meshes. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 325–334, 1999. 4

[7] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (ToG)*, 38(4):1–12, 2019. 4

[8] Bin Huang, Yangguang Li, Enze Xie, Feng Liang, Luya Wang, Mingzhu Shen, Fenggang Liu, Tianqi Wang, Ping Luo, and Jing Shao. Fast-bev: Towards real-time on-vehicle bird's-eye view perception. *arXiv preprint arXiv:2301.07870*, 2023. 3

[9] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1, 3

[10] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9223–9232, 2023. 3

[11] Alireza Javaheri, Catarina Brites, Fernando Pereira, and João Ascenso. A generalized hausdorff distance based quality metric for point cloud geometry. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2020. 6

[12] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1251–1261, 2020. 4

[13] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1042–1050, 2023. 1, 3

[14] Maik Keller, Damien Lefloch, Martin Lambers, Shahram Izadi, Tim Weyrich, and Andreas Kolb. Real-time 3d reconstruction in dynamic scenes using point-based fusion. In *2013 International Conference on 3D Vision-3DV 2013*, pages 1–8. IEEE, 2013. 1

[15] Leif Kobbelt, Swen Campagna, Jens Vorsatz, and Hans-Peter Seidel. Interactive multi-resolution modeling on arbitrary meshes. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 105–114, 1998. 4

[16] Abhinav Kumar, Yuliang Guo, Xinyu Huang, Liu Ren, and Xiaoming Liu. Seabird: Segmentation in bird's view with dice loss improves monocular 3d detection of large objects. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[17] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1477–1485, 2023. 3

[18] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9087–9098, 2023. 1, 3

[19] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 3, 4, 6, 12

[20] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022. 3

[21] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 4

[22] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 3

[23] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023. 3

[24] Zhipeng Luo, Changqing Zhou, Gongjie Zhang, and Shijian Lu. Detr4d: Direct multi-view 3d object detection with sparse attention. *arXiv preprint arXiv:2212.07849*, 2022. 3

[25] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. *arXiv preprint arXiv:2312.06741*, 2023. 1

[26] Ruihang Miao, Weizhou Liu, Mingrui Chen, Zheng Gong, Weixin Xu, Chen Hu, and Shuchang Zhou. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv preprint arXiv:2302.13540*, 2023. 3

[27] Takashi Michikawa, Takashi Kanai, Masahiro Fujita, and Hiroaki Chiyokura. Multiresolution interpolation meshes. In *Proceedings Ninth Pacific Conference on Computer Graphics and Applications. Pacific Graphics 2001*, pages 60–69. IEEE, 2001. 4

[28] Tomoyuki Mukasa, Jiu Xu, and Bjorn Stenger. 3d scene mesh from cnn depth predictions and sparse monocular slam. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 921–928, 2017. 1

[29] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*, 2022. 3

[30] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 3

[31] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 4

[32] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3437–3444. IEEE, 2023. 1

[33] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 134–144, 2019. 1

[34] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021. 1

[35] Su Sun, Cheng Zhao, Yuliang Guo, Ruoyu Wang, Xinyu Huang, Yingjie Victor Chen, and Liu Ren. Behind the veil: Enhanced indoor 3d scene reconstruction with occluded surfaces completion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 4

[36] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[37] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8406–8415, 2023. 1, 3, 6, 8

[38] Hao Wang, Nadav Schor, Ruizhen Hu, Haibin Huang, Daniel Cohen-Or, and Hui Huang. Global-to-local generative model for 3d shapes. *ACM Transactions on Graphics (TOG)*, 37(6):1–10, 2018. 4

[39] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. *arXiv preprint arXiv:2303.11926*, 2023. 1, 6

[40] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. *arXiv preprint arXiv:2303.03991*, 2023. 1, 3, 4, 6

[41] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. *arXiv preprint arXiv:2306.10013*, 2023. 1, 3

[42] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 3, 12

[43] Zitian Wang, Zehao Huang, Jiahui Fu, Naiyan Wang, and Si Liu. Object as query: Equipping any 2d object detector with 3d detection ability. *arXiv preprint arXiv:2301.02364*, 2023. 3

[44] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 1, 3

[45] Thomas Whelan, Stefan Leutenegger, Renato Salas-Moreno, Ben Glocker, and Andrew Davison. Elasticfusion: Dense slam without a pose graph. In *Robotics: Science and Systems*. Robotics: Science and Systems, 2015. 1

[46] Enze Xie, Zhiding Yu, Daquan Zhou, Jonah Philion, Anima Anandkumar, Sanja Fidler, Ping Luo, and Jose M Alvarez. M2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. *arXiv preprint arXiv:2204.05088*, 2022. 1, 3

[47] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018. 3, 4, 12, 13

[48] Yunpeng Zhang, Wenzhao Zheng, Zheng Zhu, Guan Huang, Jiwen Lu, and Jie Zhou. A simple baseline for multi-camera 3d object detection. In *AAAI Conference*, volume 37, pages 3507–3515, 2023. 3

[49] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2304.05316*, 2023. 1, 3

[50] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3

[51] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 1

Appendix

We provide in this supplementary more ablation studies and additional visualizations on AdaOcc_B that could not fit in the paper. In particular, we include (1) more details about the loss, (2) more details about the evaluation metrics, (3) the ablation study examining the impact of the number of bounding boxes in occupancy mapping, (4) the ablation study on the number of folded points for each bounding box, and (5) more occupancy visualizations for all baselines.

A. Loss Details

Semantic occupancy loss. We applied focal loss as the semantic occupancy loss. Focal loss is a typical classification loss, specialized to tackle problems such as class imbalances and hard data samples.

$$\mathcal{L}_{\text{sem}}(M) = \sum_{x_{\min}}^{x_{\max}} \sum_{y_{\min}}^{y_{\max}} \sum_{z_{\min}}^{z_{\max}} -\alpha(1-p(x,y,z))^\beta \log(p(x,y,z)),$$
(5)

where p(.) represents the predicted probability of the correct class, α and β are hyperparameters to balance well-classified and hard examples.

Object detection loss. We use a similar loss function as in DETR3D [42] for the object detection task. The object detection loss includes a focal loss and a L1 loss, for the classification and regression of the bounding boxes respectively. The focal loss is similar to Eq. 5, minimizing the discrepancy between the predicted bounding boxes classes and ground truth bounding box classes, and L1 loss minimizes the difference between N predicted bounding box parameters B_{pred} and N corresponding ground truth bounding box parameters B_{GT} as:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} |B_{\text{pred}} - B_{\text{GT}}|.$$
(6)

Surface reconstruction loss. We use the Chamfer distance [47] as the surface reconstruction loss. It is a geometric distance-based loss function used for measuring the dissimilarity between two point sets. In the context of our work, it quantifies the discrepancy between the reconstructed surface points and the ground truth points. The Chamfer distance $\mathcal{L}_{Chamfer}$ is defined as:

$$\mathcal{L}_{\text{surf}} = \sum_{\mathbf{x} \in \mathbf{X}} \min_{\mathbf{y} \in \mathbf{Y}} \|\mathbf{x} - \mathbf{y}\|_2^2 + \sum_{\mathbf{y} \in \mathbf{Y}} \min_{\mathbf{x} \in \mathbf{X}} \|\mathbf{y} - \mathbf{x}\|_2^2, \quad (7)$$

where $\mathbf{X}$ represents the reconstructed points, and $\mathbf{Y}$ denotes the ground truth points clouds.

B. Evaluation metrics details

Intersection over Union (IOU): IOU quantifies the overlap between the predicted and ground truth regions. It computes the ratio of the intersection to the union of these regions, providing an indicator of how well the prediction aligns with the actual data.

Mathematically, IOU is defined as:

$$IOU = \frac{\text{Area of Intersection}}{\text{Area of Union}}$$
(8)

Mean Intersection over Union (mIOU): mIOU is the mean value of IOU computed across multiple instances or classes. It provides a holistic measure of the method's accuracy across various categories.

Mathematically, mIOU is defined as:

$$mIOU = \frac{1}{N} \sum_{i=1}^{N} IOU_i,$$
(9)

where N is the number of instances or classes and IOU_i represents the IOU value for the ith instance or class.

However, IoU and mIoU are primarily used to measure the accuracy of object detection or segmentation models, quantifying the overlap between two regions. While they can indicate the disparity between the detected object's position and the true object's position, they do not provide detailed information about shape. Therefore, even with a high IoU, it does not guarantee the accuracy of the detected object's shape.

Hausdorff distance, on the other hand, offers a deeper metric for assessing the accuracy of object shapes. It evaluates the similarity of the evaluated object point cloud and the ground truth object point cloud by measuring the maximum pairwise distance between them. This implies that Hausdorff distance can better describe the precision of object shapes, not just their positional accuracy. Hence, Hausdorff distance is highly useful in tasks such as shape reconstruction, point cloud matching, etc., where a comprehensive consideration of object shape accuracy is necessary. It is defined as:

$$\mathcal{L}_{\text{Hausdorff}} = \max \left\{ \max_{\mathbf{x} \in X} \min_{\mathbf{y} \in Y} \|\mathbf{x} - \mathbf{y}\|, \max_{\mathbf{y} \in Y} \min_{\mathbf{x} \in X} \|\mathbf{y} - \mathbf{x}\| \right\}$$
(10)

where $\mathbf{X}$ represents the reconstructed point cloud points, and $\mathbf{Y}$ denotes the ground truth point cloud points. A detailed illustration can be referred to Fig. I.

C. Impact of the number of bounding boxes

We utilize the DeTR head introduced in [19] for the object detection task. The number of bounding boxes is a hyperparameter. We explore this parameter in our ablation study, as presented in Tab. I.
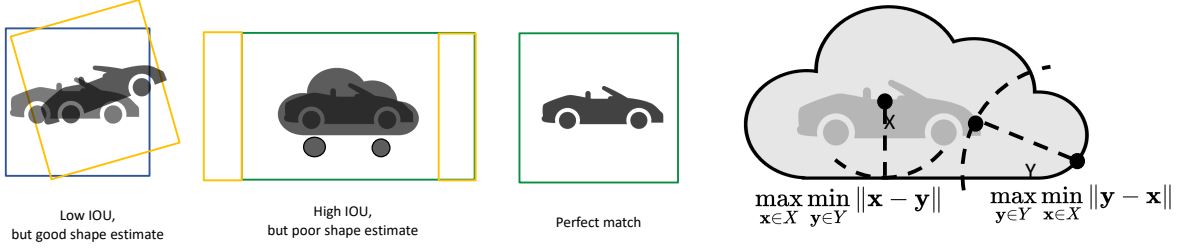
Figure I. Misaligned boxes lead to misalignment of the reconstructed objects. Moreover, we also demonstrate how Hausdorff distance works.

Table I. **Ablation study on the number of bounding boxes for folding.** We adjust the number of bounding boxes from 0 to 40, with 0 representing the original BEVFormer setup.

# of box	10	20	30	40	0
IOU	**0.313**	0.312	0.312	0.310	0.309
mIOU	**0.1590**	0.1532	0.1488	0.1475	0.1540
Time(s)	0.1201	0.1298	0.1352	0.1603	**0.1109**

D. Study on different methods of Compressing 3D into 2D

We investigate various methods of compressing 3D features into 2D, which is required specifically for AdaOcc-tasks based on the CONet backbone, such as object detection and folding. The size of the 3D features varies depending on the size of the bounding boxes. Thus, a fixed attention-based weighted aggregation does not work as the input dimension is not fixed. Therefore, we selected four different aggregation methods: max pooling, average pooling, global mean (taking the average over all dimensions), and global max (taking the max over all dimensions). Note that we do not consider summation to be a viable method, as the varying size of 3D features would make the summed 2D features heavily dependent on the size of the 3D features, which is undesirable. It turns out that the max pooling layer performs best, which is similar to the aggregation method used in PointNet, as shown in Tab. II.

Table II. **Ablation study on different methods of Compressing 3D into 2D** We investigate various methods of compressing 3D features into 2D, which is required specifically for AdaOcc tasks based on the BEVFormer backbone. The methods examined include maxpooling, avgpooling, global-mean, and global-max. In this ablation, we set the train grid size to 0.8m.

Method	Maxpooling	Avgpooling	Global-mean	Global-max
IOU	0.090	0.087	0.088	0.088
mIOU	0.053	0.051	0.049	0.047

E. Study on the number of folded points per Box

We employ FoldingNet [47] to reconstruct the surface of objects. While the number of folded points for training remains fixed, the number of folded points for testing can be adjusted based on resolution requirements. We conduct an ablation study to analyze how the number of folded points in testing affects IOU and mIOU results, shown in Table. III.

Table III. **Ablation study on the number of folded points per box.** We vary the number of points from 900 to 40,000. In our previous experiment setup, we opt for 2500 points as it offers the best balance between performance and evaluation time.

Fold size	900	2500	10000	40000
IOU	0.312	**0.313**	**0.313**	**0.313**
mIOU	0.1530	0.1532	0.1537	**0.1541**
Time(s)	0.1414	**0.14042**	0.1419	0.1444

F. More occupancy visualization

Additional occupancy maps for various scenes are displayed. AdaOcc demonstrates improved separation of the bounding boxes.
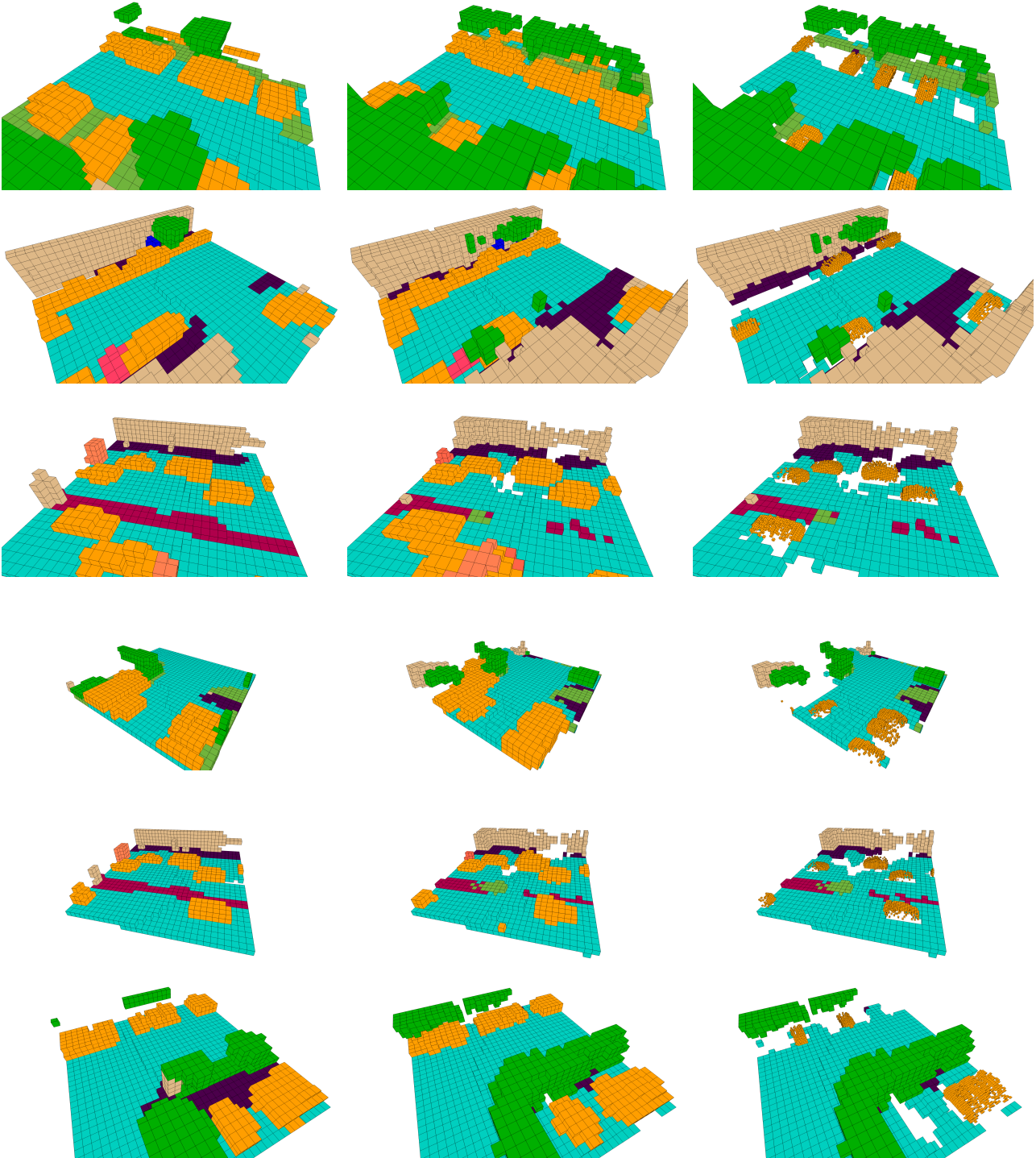
Figure II. **Visualization on BEVFormer vs CONet vs. AdaOcc (We visualize the pointcloud in a voxelized format)**.