Efficient Occupancy Prediction with Instance-level Attention

Sungjin Park*
Department of Automotive Engineering
Hanyang University
Seoul, Republic of Korea
shihtzu333@hanyang.ac.kr

Jaeha Song*
Department of Automotive Engineering
Hanyang University
Seoul, Republic of Korea
archiiive99@hanyang.ac.kr

Soonmin Hwang†
Department of Automotive Engineering
Hanyang University
Seoul, Republic of Korea
soonminh@hanyang.ac.kr

***Abstract*—Occupancy prediction is a critical task in autonomous driving, enabling better understanding of 3D environments for downstream tasks. Previous methods often rely on dense back-projection methods to extract 3D features from 2D images by distributing information across all voxels. While effective, these approaches are computationally expensive and inefficient due to the dense nature of 3D voxel representations. Inspired by recent works, we address this challenge by instance-level attention that utilizes representative queries for groups of voxels, reducing computational cost while maintaining competitive performance. By applying attention mechanisms to this instance-level representation, we achieve an mIoU score of 35.26 with a latency of 0.04s on the Occ3D dataset in RTX4090. These results demonstrate that focusing on instance-level representations provides an efficient and practical solution for real-time occupancy prediction tasks.**

***Index Terms*—Occupancy Prediction, View Transformation, Occupancy Network**

I. Introduction

With the rapid advancement of Autonomous Driving (AD) technology, recent developments have pushed the boundaries toward End-to-End AD, which integrates perception and planning into a single-shot framework. In this context, recent studies [1] have highlighted the critical role of the occupancy prediction task for the next generation of AD systems.

Conventional methods for 3D occupancy prediction have relied on either backward-projection [2] or forward-projection [3], [4] approaches, each with its own limitations. Former methods process every voxel within the field of view, resulting in high computational costs that affect real-time deployment. Later approaches, on the other hand, project 2D image pixels into 3D space but often struggle to fully capture the 3D scene context, particularly in ambiguous or occluded regions, as well as due to the inherent uncertainty in depth prediction.

To address these challenges, we propose an instance-level attention with instance queries and 2D representation map, aiming to achieve a balance between performance and computational efficiency. By leveraging meaningful instance-level features extracted from 2D images, our method reduces reliance on dense 3D representations while enhancing scene comprehension, and compressing the 3D feature map into a 2D representation, we significantly reduce computational complexity while preserving spatial information. Experimental results on the Occ3D-nuScenes dataset demonstrate that our approach achieves competitive performance compared to other models, while significantly reducing latency. These results highlight the potential of instance-based reasoning as an efficient and scalable alternative for 3D occupancy prediction in AD.

II. Related Works

A. View Transformation for Occupancy Prediction

To handle 2D images in 3D space, view transformation from 2D to 3D is essential. Forward-projection, introduced for BEV perception in Lift-Splat-Shoot [3], estimates the depth of images and transforms them into 3D space using camera parameters. While this method is simple and computationally efficient, it operates on 2D pixels rather than 3D voxels, resulting in sparsity in the 3D map and ambiguity in depth estimation. BEVDepth [4] partially addresses this depth ambiguity by incorporating auxiliary depth loss during training.

On the other hand, back-projection methods process information based on 3D voxels, effectively addressing the sparsity issue inherent in forward-projection. However, this approach comes with the drawback of significantly higher computational costs. Additionally, since back-projection maps 3D voxels back to 2D image features, it introduces another form of ambiguity, as information can stretch along the ray direction, degrades precision in localization.

To overcome these limitations, FB-OCC [5] proposed a hybrid approach that first constructs a 3D map using forward-projection and then refines the mapped points through backward-projection. This method leverages the strengths of both approaches, achieving high performance by combining the efficiency of forward-projection with the detailed refinement capabilities of backward-projection. However, despite its strong performance, FB-OCC still falls short of real-time applicability due to its computational demands.

Recently, sparse back-projection methods [6]–[8] have been introduced to further reduce computational complexity. Instead of processing all 3D voxels in the map, these methods select only a few key voxels and iteratively back-project their surrounding offset voxels. By focusing computations on a sparse

*Equal contribution.
†Corresponding author.

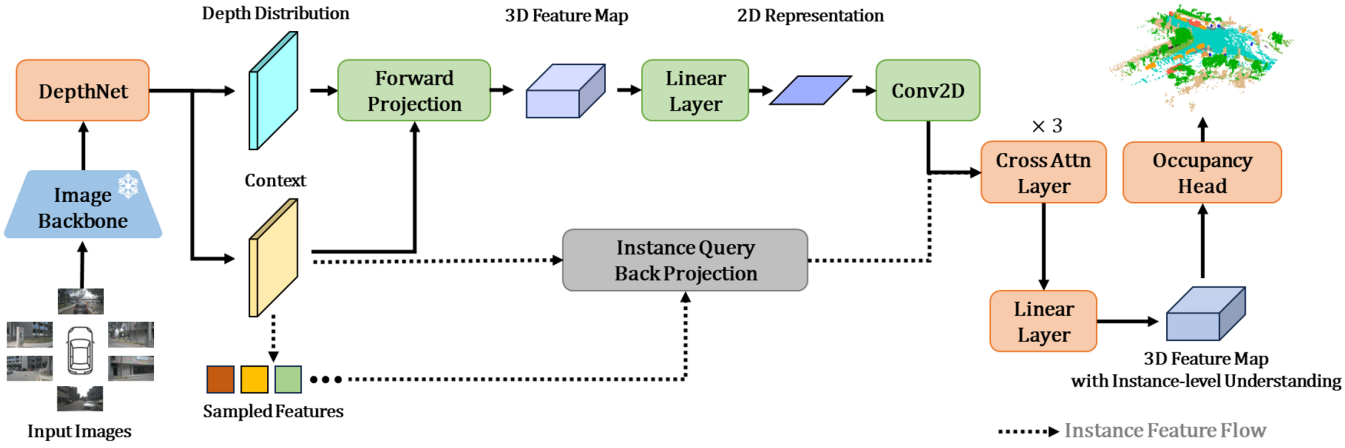

Fig. 1. **Overall Architecture of the Proposed Network.** The pipeline consists of a backbone for extracting 2D image features and a depth network for predicting depth distribution and encoding context. Based on FB-OCC [5], the architecture includes two complementary paths: one for constructing the base feature map and another for extracting instance-level features for queries. Cross Attention Layer between these paths enables the feature map to incorporate instance-level scene understanding.

subset of the 3D map, these approaches significantly reduce computational overhead while improving performance.

B. Instance-level Scene Understanding with Attention

In existing occupancy prediction pipelines, view transformation and subsequent processes are voxel-centric, focusing on reconstructing the scene at a voxel-level. While effective for many tasks, this approach often lacks sufficient instance-level scene understanding, which is critical for accurately capturing object shapes in 3D occupancy prediction.

Symphonies [9] addresses this gap by introducing learnable instance queries, with multiple rounds of attention between image features, instance queries, and 3D voxel map. This approach successfully enhances instance-level understanding, leading to notable performance improvements. However, it comes with certain drawbacks. The use of learnable instance queries for every scene can result in inefficiencies, as some queries may not correspond to any actual objects in a given scene. Additionally, the computational cost increases significantly due to the repeated attention operations required to refine the instance-level representation.

Despite these limitations, Symphonies demonstrated the importance of instance-level scene understanding in improving performance, highlighting its necessity for achieving more accurate and comprehensive 3D occupancy prediction.

III. Efficient Occupancy Prediction

This section provides a detailed explanation of our proposed network. Based on FB-OCC [5], our network focuses on utilizing instances-level query with attention while incorporating the 2D representation approach from FlashOcc [10] to improve computational efficiency.

A. Overall Architecture

As illustrated in Fig. 1, the overall architecture of our proposed network begins with a forward projection that projects features from backbone. The resulting 3D feature map is then flattened into a 2D representation. Independently, the model identifies meaningful points in each image and extracts instance-level features using deformable attention with learnable queries to form instance queries. These instance queries are used as values in an transformer layer, with the flattened 2D representation as queries, enabling the network to achieve an instance-level scene understanding. Finally, the processed features are restored along the height dimension to reconstruct a 3D representation, which is passed to the occupancy head for the final prediction.

B. View Transformation

To construct the 3D feature map, we employ the forward projection methods from Lift-Splat-Shoot [3]. As this 3D feature map $\mathbf{F}^{3D} \in \mathbb{R}^{D\times W\times H\times C}$ serves as queries in the subsequent transformer layer, it is flattened into 2D representation $\mathbf{F}^{2D} \in \mathbb{R}^{D\times W\times C'}$ to reduce computational cost. Inspired by FlashOcc [10], it compresses the height with the channels from $H \cdot C$ to C' with a simple linear layer followed by 3×3 2D convolution with LayerNorm(LN), while preserving the spatial dimensions $D \times W$.

C. Feature Aggregation for Instance Queries

Independent to the view transformation, instead of using every voxel in visible area as attention sequences in FB-OCC [5], we identify several meaningful points in each image based on the norm value of features extracted from the backbone, represented as $\mathbf{P} \in \mathbb{R}^{N_{\text{cam}}\times N_{\text{points}}\times 2}$. Using these points as reference points, Deformable Attention [11] is applied to features of these points, $\mathbf{F}^{inst} \in \mathbb{R}^{N_{\text{cam}}\times N_{\text{queries}}\times C}$, to integrate features at the instance level.

Deformable Attention denoted as DeformAttn, is based on Attention denoted as Attn, which can be expressed as:

$$\text{Attn}(Q, K, V) = \sum_i \text{softmax}\left(\frac{QK_i^\top}{\sqrt{d_k}}\right) V_i$$

Here, $Q \in \mathbb{R}^{N_q \times d_k}$ represents the query matrix, and $K, V \in \mathbb{R}^{N_k \times d_k}$ are the key and value matrices, representing features. Then d_k is the dimensionality of the key and query. Deformable Attention adaptively apply the above attention mechanism to queries q, with their corresponding value, 2D image feauters F^{2D} in our case, and its position referred to as reference points p_q. This can be expressed as:

$$\text{DeformAttn}(q, p_q, F^{2D}) = \sum_{k=1}^{K} A_{qk} W F^{2D}(p_q + \Delta p_{qk})$$

Here, K represents the number of sampling points, and A_{qk} is attention weight for query q and sampled points, which are reference point p_q and it offsets Δp_{qk}. W represents feature weight. By limiting attention to only the designated sampled points, Deformable Attention significantly reduces computational cost compared to base attention mechanism.

We additionally encode camera parameters into these instance features using the linear layer, resulting in an output represented as $\mathbf{F}^{inst} \in \mathbb{R}^{N_{\text{cam}} \times N_{\text{queries}} \times C'}$, enabling the model to account for variations across different views.

D. Instance-level Attention

The model then performs Spatial Cross Attention(SCA) layer, leveraging the 2D representation from Sec III-B as the query Q and the instance features as the key K and value V for the Attn in Sec III-C. This enables the model to achieve an instance-level understanding of the map.

Subsequently, a simple linear layer with LN is employed for height reconstruction similar to the height compression in Sec III-B, restoring the height dimension from C' to $H \cdot C$. The refined features are then passed through 3D convolutions in the occupancy head to produce the final predictions.

E. Losses

For loss function, we adopt the Scene-Class Affinity Loss L_{scal} from MonoScene [12], which simultaneously optimizes precision, recall, and specificity. This is applied to both geometric(L_{sca}^{geo}) and semantic(L_{sca}^{sem}) predictions, with focal lossL_{focal} addressing the class imbalance by emphasizing harder-to-classify samples. This approach enhances the model's robustness, particularly for rare classes and ambiguous regions in the scene.

Additionally, since our method is based on forward projection, the depth cross-entropy loss(L_{depth}) from BEVDepth [4] is incorporated as an auxiliary loss to improve localization accuracy. The overall loss function is expressed as:

$$L = L_{scal}^{geo} + L_{scal}^{sem} + L_{focal} + L_{depth}$$

IV. Experiments

A. Dataset and Metric

We evaluate our model on the Occ3D-nuScenes [13] dataset, which is derived from the nuScenes [17] dataset. The nuScenes dataset contains large-scale multimodal data collected from 6 surround cameras, 1 LiDAR, and 5 radars. However, in our setting, we use data from only the 6 surround cameras.

To evaluate the performance of our model, we use the mean Intersection over Union (mIoU) metric, as it is a widely recognized standard for measuring the performance of occupancy prediction. This metric ensures a fair and consistent comparison with previous research.

B. Implementation Details

Our implementation is built upon FB-OCC [5], utilizing ResNet-50 [18] as the backbone. The Deformable Attention for instance features described in Sec III-C consists of a single layer, while the Spatial Cross Attention (SCA) module in Sec III-D comprises three layers. We sampled 100 features from the context per view, 600 sampled features in total.

For training, we apply the AdamW [19] optimizer with a batch size of 10. The initial learning rate is set to 3×10^{-4} and follows a linear decay policy. Our model is trained for 30 epochs. Latency is measured on an NVIDIA RTX 4090 GPU using PyTorch in fp32 precision.

C. Results

In Table.I, we compare our model with other occupancy prediction models on the Occ3D-nuScenes validation set. To ensure a fair comparison, we use results from other models that also exclude history frames, as our model operates on a single time step. Using input images of size 256×704 with a ResNet-50 [18] backbone, our model does not surpass the highest performance achieved by FB-OCC [5], which serves as our base model. However, it demonstrates competitive performance compared to other models.

When examining specific object categories, our model shows minimal performance loss for distinct objects such as cars and bicycles, indicating that it effectively captures these foreground features. However, the performance drop is more significant for background categories like sidewalks and terrain. We hypothesize that this discrepancy arises because background regions are sampled less frequently compared to object features, leading to reduced representation and poorer predictions for these areas.

In terms of latency, FB-OCC [5] relies on dense attention, which leads to significant latency degradation. FlashOcc [10], on the other hand, employs a 2D representation, achieving faster latency but at the cost of notable drop in performance. Our model, combining FB-OCC's architecture [5] with a reduced attention sequence and FlashOcc's 2D representation [10], strikes a balance between latency and performance, outperforming FlashOcc [10] while being faster than FB-OCC [5].

Focusing on instance-level scene understanding with reduced computational overhead, our approach demonstrates

TABLE I
OCCUPANCY PREDICTION PERFORMANCE ON THE OCC3D-NUSCENES DATASET (SINGLE-FRAME VALIDATION)

Method	Resolution	Backbone	**mIoU↑**	**others**	**barrier**	**bicycle**	**bus**	**car**	**cons. veh.**	**motorcycle**	**pedestrian**	**traffic cone**	**trailer**	**truck**	**drive. surf.**	**other flat**	**sidewalk**	**terrain**	**manmade**	**vegetation**	**Latency(ms)↓**
MonoScene [12]	928 × 1600	R101	6.06	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	-
CTF-Occ [13]	928 × 1600	R101	28.53	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	18.00	-
TPVFormer [14]	928 × 1600	R101	27.83	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	-
FlashOcc(TRT-FP16) [10]	256 × 704	R50	32.08	6.74	37.65	10.26	39.55	44.36	14.88	13.4	15.79	15.38	27.44	31.73	78.82	37.98	48.7	52.5	37.89	32.24	7.32
OccFormer [15]	256 × 704	R50	20.40	6.62	32.57	13.13	20.37	37.12	5.04	14.02	21.01	16.96	9.34	20.64	40.89	27.02	27.43	18.65	18.78	16.90	311
BEVDetOcc [16]	256 × 704	R50	31.64	6.65	36.97	8.33	38.69	44.46	15.21	13.67	16.39	15.27	27.11	31.04	78.70	36.45	48.27	51.68	36.82	32.09	412
FB-OCC [5]	256 × 704	R50	**37.39**	**12.17**	**44.83**	**25.73**	**42.61**	**47.97**	**23.16**	**25.17**	**25.77**	**26.72**	**31.31**	**34.89**	**78.83**	**41.42**	**49.06**	**52.22**	**39.07**	**34.61**	94.2
Ours	256 × 704	R50	35.26	11.08	42.79	23.02	41.35	47.01	22.1	24.51	25.21	25.01	28.78	34.36	76.24	37.3	43.84	47.34	36.91	32.51	**31.8**

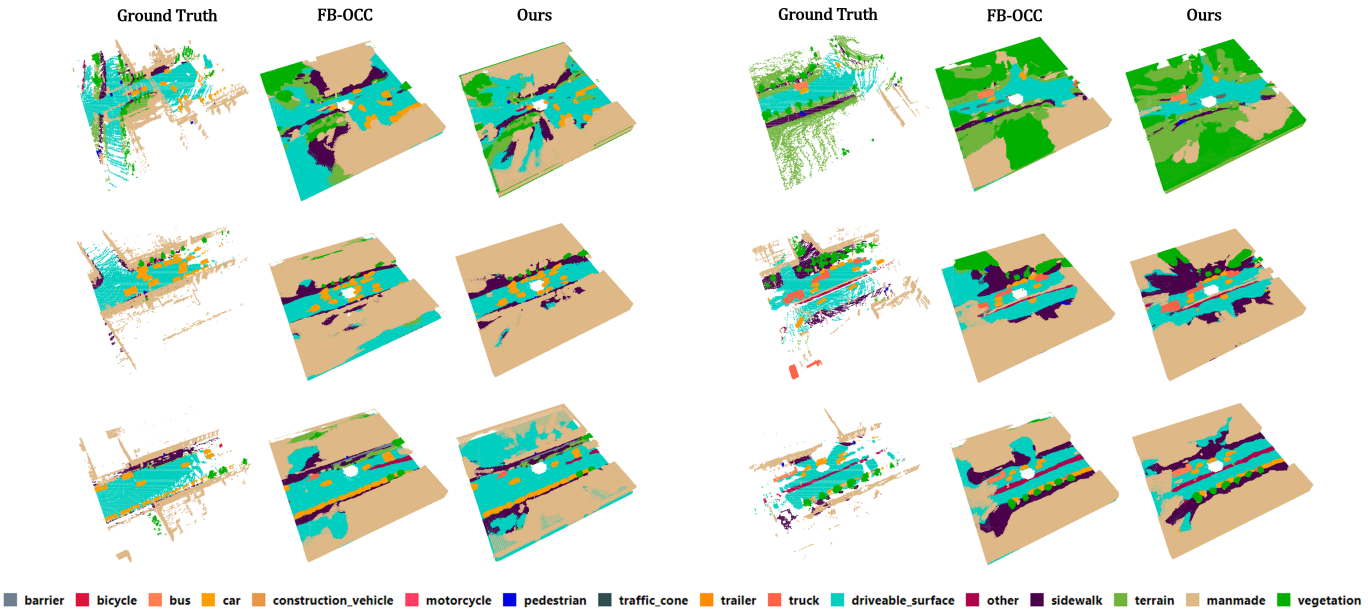

Fig. 2. **Qualitative results on the Occ3D-nuScenes validation set.** We illustrate the occupancy prediction outcomes, each color represents a specific class, as indicated in the accompanying legend. To ensure consistency with the prediction process, we masked the areas outside the field of view, as they are not covered by surround cameras.

the ability to effectively capture the entire scene context. Furthermore, our results suggest that dense attention may not always be essential for view transformation, providing a more efficient and practical alternative.

D. Visualization

In Fig. 2, we present the qualitative results of our occupancy prediction compared to the ground truth. As our method relies solely on single-frame input without utilizing temporal fusion, it tends to reconstruct occluded regions in a somewhat random manner due to the exclusion of occluded areas from the loss calculation during training. When compared to FB-OCC [5], our method demonstrates a similar level of reconstruction for objects, indicating its ability to effectively capture the presence of objects.

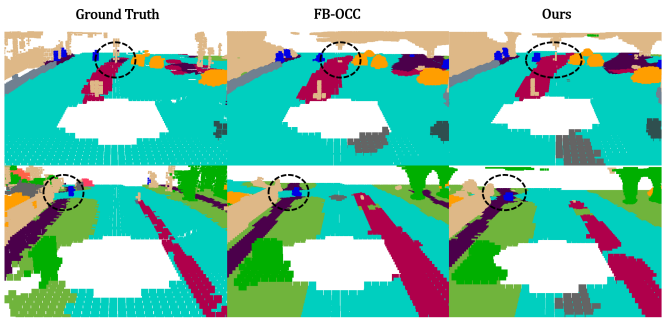

Fig. 3. **Qualitative comparison of background between FB-OCC and our method.** While both models demonstrate similar capabilities in object reconstruction, our method struggles to capture background details as effectively as FB-OCC [5].

However, as shown in Fig. 3, our method exhibits qualitatively inferior performance in reconstructing background regions. This discrepancy highlights a limitation in our approach, as background features are sampled less frequently during training. Furthermore, since the ground truth does not provide complete object shapes, our model struggles to fully restore these objects to their complete form, further emphasizing the challenges of precise reconstruction.

V. Conclusion

In this paper, we propose utilizing instance-level attention for scene understanding that achieves a balance between performance and latency. By addressing the computational burden of dense backward-projection and the sparsity of forward projection, our method demonstrates modest yet meaningful improvements in performance while reducing latency. Experiment results validate the effectiveness of this approach for the occupancy prediction.

A. Limitations

Our method currently samples features based solely on norm scores, which limits its adaptability. This approach tends to focus on objects while frequently underrepresenting background regions. Furthermore, important instances with low norm scores, such as small or partially occluded objects, may also be overlooked. To address this limitation, future work could explore adaptive feature selection methods. For instance, leveraging 2D image segmentation ground truth could allow the model to explicitly learn and focus on meaningful segment regions. Specifically, an auxiliary segmentation loss could be introduced to evaluate whether the sampled features accurately align with the semantic segmentation regions. This loss would penalize the model when sampled features fail to correspond to semantically important areas, ensuring the selection process prioritizes relevant regions. By aligning the sampled features with 2D segmentation masks, the model could enhance its adaptability and robustness, improving its ability to detect small or partially occluded instances even under challenging conditions.

Additionally, while temporal attention is critical for addressing occlusion and localization ambiguity, our current approach does not yet incorporate temporal information into predictions. Existing fusion methods typically utilize entire 3D feature sets from past frames, which can be computationally expensive. Extend to our approach, we believe that temporal fusion could be achieved more effectively and efficiently by selectively sampling key features across time steps. For instance, integrating these sampled key features along the temporal axis with motion-based feature alignment techniques could improve temporal consistency while reducing computational overhead.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00409492).

References

[1] X. Weng, B. Ivanovic, Y. Wang, Y. Wang, and M. Pavone, "Paradrive: Parallelized architecture for real-time autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15449–15458, 2024.

[2] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*, pp. 1–18, Springer, 2022.

[3] J. Philion and S. 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*, pp. 194–210, Springer, 2020.

[4] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 1477–1485, 2023.

[5] Z. Li, Z. Yu, D. Austin, M. Fang, S. Lan, J. Kautz, and J. M. Alvarez, "Fb-occ: 3d occupancy prediction based on forward-backward view transformation," *arXiv preprint arXiv:2307.01492*, 2023.

[6] X. Lin, T. Lin, Z. Pei, L. Huang, and Z. Su, "Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion," *arXiv preprint arXiv:2211.10581*, 2022.

[7] H. Liu, H. Wang, Y. Chen, Z. Yang, J. Zeng, L. Chen, and L. Wang, "Fully sparse 3d panoptic occupancy prediction," *arXiv preprint arXiv:2312.17118*, 2023.

[8] L. Chambon, E. Zablocki, M. Chen, F. Bartoccioni, P. Pérez, and M. Cord, "Pointbev: A sparse approach for bev predictions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15195–15204, 2024.

[9] H. Jiang, T. Cheng, N. Gao, H. Zhang, T. Lin, W. Liu, and X. Wang, "Symphonize 3d semantic scene completion with contextual instance queries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20258–20267, 2024.

[10] Z. Yu, C. Shu, J. Deng, K. Lu, Z. Liu, J. Yu, D. Yang, H. Li, and Y. Chen, "Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin," *arXiv preprint arXiv:2311.12058*, 2023.

[11] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2020.

[12] A.-Q. Cao and R. De Charette, "Monoscene: Monocular 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3991–4001, 2022.

[13] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9223–9232, 2023.

[15] Y. Zhang, Z. Zhu, and D. Du, "Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9433–9443, 2023.

[16] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[17] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.