Improving 3D Semantic Occupancy Prediction using Spatiotemporal Transformers

I. Introduction

Accurate and comprehensive 3D scene understanding and reasoning are crucial for the development of autonomous driving systems. In recent years, there has been significant interest in vision-centric 3D perception [5, 28, 23, 60, 52, 66] as a promising alternative to LiDAR-based methods [24, 71, 44, 3, 47] in autonomous driving academia. While LiDAR-based methods, which rely on explicit depth measurements, have demonstrated leading performance on public datasets [16, 4, 2, 50, 20], vision-based approaches offer distinct advantages in terms of cost-effectiveness and the ability to detect long-range distance objects. Furthermore, vision-based methods excel in identifying road elements such as traffic lights and road signs, which is a valuable feature compared to their LiDAR-based counterparts. This growing attention towards vision-based 3D perception in the context of autonomous driving reflects its potential to enhance the capabilities and efficiency of autonomous vehicles in real-world scenarios.

In the realm of vision-centric perception for autonomous driving systems, the utilization of multiple cameras has gained prominence in capturing both spatial and temporal cues from 2D RGB images. While monocular methods [5, 66, 43, 46] offer a straightforward solution, they tend to process individual camera views independently, limiting their ability to capture and leverage information across multiple cameras. In contrast, multi-camera methods [13, 28, 23, 60, 52] have emerged as a compelling unified alternative, demonstrating progress in the realm of 2D-to-3D transformation. By constructing comprehensive representations of the surrounding scene, these methods enable various applications, including 3D object detection, semantic occupancy prediction, and semantic scene completion.

The task of lifting 2D perspective observations into 3D space is an ill-posed problem due to the loss of depth information in the formation of 2D images. However, with the utilization of strong prior information, this task remains tractable. The effectiveness of 3D scene understanding heavily relies on the representation of the 3D environment, as illustrated in Figure 1. Traditional approaches [56, 46] involve dividing the 3D space into voxels and assigning each voxel a vector to denote its status. However, this representation proves computationally expensive. Alternatively, Bird's Eye View (BEV) representation, which disregards height and focuses on the top-down view, offers a more efficient solution. While BEV-based methods [28, 22, 13, 27, 34, 32, 37] perform remarkably well in 3D object detection, they struggle to encode the 3D structure of objects, thus hindering performance in 3D semantic occupancy prediction. To overcome this limitation, TPVFormer [23] introduces a Tri-Perspective View (TPV) representation, incorporating two additional perpendicular planes. This hybrid explicit-implicit representation [8, 7] aims to capture both the efficiency of BEV and the ability to encode the 3D structure of objects.

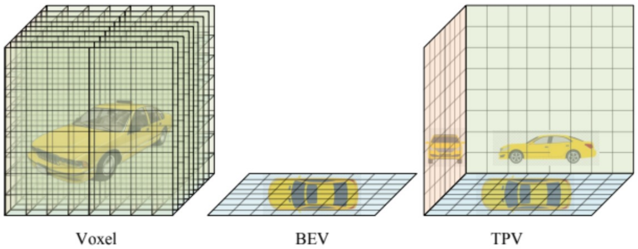

Fig. 1. Comparison of Voxel, BEV and TPV [23] latent vector fields used to represent 3D scenes. While BEV is more efficient than the Voxel representation, it discards the height information and cannot provide a holistic understanding of a 3D scene. TPV is an approach to increase efficiency without compromising the accuracy of representation. But, TPV is susceptible to the deficiency of fine-grained semantic information.

The Semantic Scene Completion (SSC) task was initially introduced in the SSCNet paper [49] and gained further prominence with the advent of SemanticKITTI [2], which provided an official dataset and competition track. More recently, another related task called Semantic Occupancy Prediction (SOP) has emerged, albeit with slight distinctions. Both SSC and SOP share the common objective of predicting the occupancy status and semantic class of a voxel at a given spatial location. However, there exist subtle differences between the two. Firstly, SSC utilizes partial 3D data obtained from LiDAR or other active depth sensors, thus justifying its name as a ”completion” of a 3D semantic scene. On the other hand, SOP relies on 2D images, potentially from multiple cameras and frames, as its input. Additionally, while SSC typically focuses on static scenes, SOP is designed to handle dynamic objects as well.

II. Related Work

A. Latent representations for 3D perception

In the context of vision-centric 3D perception, a disparity exists between the inputs, which comprise 2D images, and the desired outputs, which entail a representation of a 3D space. To bridge this disparity, researchers have delved into vision-based approaches that transform 2D perspective image features into their 3D counterparts [28, 37, 22, 27]. These vision-based 2D to 3D transformation methods subsequently enable a range of downstream tasks, including 3D object detection [48, 54, 58, 28, 30, 22], semantic map construction [37, 69, 42, 36], and motion prediction [71, 1, 21].

LSS [37] and its subsequent works [27, 21, 67, 41] employ a prediction mechanism that estimates the depth distribution at the pixel level, facilitating the projection of image features into 3D points. These 3D points are subsequently voxelized to generate 3D perspective features. Voxel-based scene representation techniques are employed in this context, involving the discretization of the 3D space into voxels and the characterization of each voxel through a vector feature. Voxel-based methods find application in various tasks, including LiDAR segmentation [29, 51, 12, 64, 63] and 3D scene completion [44, 5, 9, 25, 61].

In recent studies, BEVFormer [28] and TPVFormer [23] have incorporated deformable attention mechanisms [72, 53] to enhance the fusion of Bird's Eye View (BEV) and Tri-Perspective View (TPV) queries with corresponding image features. TPVFormer introduces a tri-perspective view approach for predicting 3D occupancy. However, it should be noted that the output of TPVFormer exhibits sparsity due to the utilization of LiDAR supervision.

The compression of 3D scenes into 2D ground planes has proven to strike a remarkable balance between performance and efficiency in tasks like 3D object detection, semantic map construction, and motion prediction. This approach succeeds because these tasks often necessitate predictions in the form of rigid bounding boxes or Bird's Eye View (BEV) representations. However, it is important to acknowledge that condensed BEV feature maps alone cannot restore a comprehensive understanding of real-world 3D scenes. Consequently, the need arises for a more detailed and nuanced 3D representation to address the challenge of 3D semantic occupancy comprehension.

B. Vision-based 3D perception

Leveraging cameras as input offers several advantages, including cost-effectiveness, the ability to detect long-range distance objects, and the identification of vision-based road elements such as traffic lights and stop lines.

Vision-based approaches for 3D semantic occupancy prediction involve processing a T-frame historical sequence of N surround view camera images as input, utilizing known camera intrinsic and extrinsic parameters. The objective is to estimate the state of each voxel within a 3D scene. Typically, the voxel state is represented in a 2-dimensional format, indicating whether the voxel is ”occupied,” ”free,” or ”unobserved,” with semantic labels assigned to ”occupied” voxels to specify their category. Alternatively, a label such as ”GO” may be used to indicate general/unknown objects.

There exists a range of successful LiDAR-based methods for 3D object detection [70, 24, 62, 57, 39, 40, 14, 45] and 3D perception [24, 65, 38, 47, 71, 44], which have been extensively evaluated and benchmarked using public datasets [17, 4, 50, 2]. Recent advancements have focused on the development of LiDAR-camera fusion methods for 3D object detection [38, 10, 32], as well as vision-based 3D perception methods that incorporate information from multiple views to estimate depth in the surrounding environment [19, 59]. Additionally, vision-based approaches have demonstrated competitive performance in tasks such as 3D object detection [28, 30, 32, 22, 27], and semantic map construction [1, 21, 67]. These advancements have sparked rapid progress in vision-based 3D perception, challenging the dominance of LiDAR-based methods.

Vision-based 3D surround perception methods face the challenge of lacking direct geometric input, necessitating the inference of 3D scene geometry based on semantic cues. These methods can be categorized into depth-based methods and other approaches. Depth-based methods explicitly predict depth maps from image inputs to extract 3D geometric information [32, 37, 22, 27, 67, 41, 56, 33, 35, 32]. A commonly adopted pipeline involves predicting categorical depth distributions and utilizing them to project semantic features into 3D space [37]. On the other hand, other methods implicitly learn 3D features without explicitly generating depth maps [54, 28, 58, 30, 55, 11]. For instance, BEVFormer [58] incorporates cross attention mechanisms to progressively refine BEV grid features using 2D image features.

C. Semantic Scene Completion and 3D semantic occupancy prediction

3D Semantic Occupancy Prediction (SOP) and Semantic Scene Completion (SSC) are closely interconnected tasks. The objective of 3D semantic occupancy prediction is to reconstruct the detailed geometry and semantic information of the surrounding 3D environment. When the input consists of LiDAR point cloud data, this process is referred to as LiDAR segmentation and encompasses both sparse training-testing for 3D SSC and dense training-testing scenarios. The pioneering work on SSC was introduced by SSCNet [49], which jointly addresses the inference of geometry and semantics in the scene.

The distinction between 3D Semantic Occupancy Prediction (SOP) and Semantic Scene Completion (SSC) lies in their respective focuses. SSC primarily concerns itself with inferring occluded regions based on the visible parts of the scene, whereas SOP does not aim to estimate the invisible regions. Additionally, SSC typically operates in the context of static scenes, whereas SOP extends its applicability to dynamic scenes.

MonoScene [5] is a pioneering work in vision-based 3D perception, specifically focusing on Semantic Scene Completion (SSC). It introduces the first monocular framework for SSC, enabling the reconstruction of outdoor scenes using RGB inputs alone. The approach employed by MonoScene involves initially constructing a 3D feature through sight projection, followed by processing it using a classical 3D UNet. Building upon this foundation, TPVFormer [23] extends the concept to multi-camera setups, enabling 3D semantic occupancy prediction using multiple cameras.

The approach employed by MonoScene [5], which relies on 3D convolutions, has certain limitations. One such limitation is that it reasons semantics within a relatively fixed receptive field, which may not account for the varying distribution patterns of different semantic classes. Additionally, the spatial invariance of 3D convolution does not effectively handle the sparse and discontinuous 3D features generated by state-of-the-art image-to-3D transformation techniques [37, 22, 27]. Moreover, 3D convolution can require a substantial number of parameters, resulting in inefficiency for long-range methods.

Following the successful introduction of vision transformers [15, 31] in various vision tasks [47, 31, 6, 68, 18, 26], the adoption of this architecture has become prevalent. The notable achievements attained with vision transformers have served as a motivation for leveraging the attention mechanism in the construction of encoder-decoder networks for 3D semantic occupancy prediction.

References

[1] Adil Kaan Akan and Fatma Güney. "StretchBEV: Stretching Future Instance Prediction Spatially and Temporally". In: 2022.

[2] Jens Behley et al. "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences". In: (Apr. 2019).

[3] Alexandre Boulch et al. "ALSO: Automotive Lidar Self-supervision by Occupancy estimation". In: 2023.

[4] Holger Caesar et al. "nuScenes: A multimodal dataset for autonomous driving". In: (Mar. 2019).

[5] Anh-Quan Cao and Raoul de Charette. "MonoScene: Monocular 3D Semantic Scene Completion". In: *CVPR*. 2022.

[6] Nicolas Carion et al. "End-to-End Object Detection with Transformers". In: 2020.

[7] Eric R. Chan et al. "Efficient Geometry-aware 3D Generative Adversarial Networks". In: 2022.

[8] Anpei Chen et al. "TensoRF: Tensorial Radiance Fields". In: 2022.

[9] Xiaokang Chen et al. "3D Sketch-aware Semantic Scene Completion via Semi-supervised Structure Prior". In: 2020.

[10] Xuanyao Chen et al. "FUTR3D: A unified sensor fusion framework for 3D detection". In: (Mar. 2022).

[11] Yongjian Chen et al. "Monopair: Monocular 3d object detection using pairwise spatial relationships". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 12093–12102.

[12] Ran Cheng et al. "(AF)2-S3Net: Attentive Feature Fusion with Adaptive Feature Selection for Sparse Semantic Segmentation Network". In: 2021.

[13] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. "NEAT: Neural Attention Fields for End-to-End Autonomous Driving". In: *International Conference on Computer Vision (ICCV)*. 2021.

[14] Zhipeng Ding, Xu Han, and Marc Niethammer. "VoteNet: A deep learning label fusion method for multi-atlas segmentation". In: (Apr. 2019).

[15] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: 2021.

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? The KITTI vision benchmark suite". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012, pp. 3354–3361.

[18] Rohit Girdhar et al. "Video Action Transformer Network". In: 2019.

[19] Vitor Guizilini et al. "Full Surround Monodepth from Multiple Cameras". In: 2021.

[20] Timo Hackel et al. "Semantic3D.net: A new Large-scale Point Cloud Classification Benchmark". In: 2017.

[21] Anthony Hu et al. "FIERY: Future Instance Prediction in Bird's-Eye View from Surround Monocular Cameras". In: 2021.

[22] Junjie Huang et al. "BEVDet: High-performance multi-camera 3D object detection in Bird-Eye-View". In: (Dec. 2021).

[23] Yuanhui Huang et al. "Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction". In: 2023.

[24] Alex H. Lang et al. "PointPillars: Fast Encoders for Object Detection from Point Clouds". In: 2019.

[25] Jie Li et al. "Anisotropic Convolutional Networks for 3D Semantic Scene Completion". In: 2020.

[26] Yanghao Li et al. "Exploring Plain Vision Transformer Backbones for Object Detection". In: 2022.

[27] Yinhao Li et al. “BEVDepth: Acquisition of reliable depth for multi-view 3D object detection”. In: (June 2022).

[28] Zhiqi Li et al. “BEVFormer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers”. In: (Mar. 2022).

[29] Venice Erin Liong et al. “AMVNet: Assertion-based Multi-View Fusion Network for LiDAR Semantic Segmentation”. In: 2020.

[30] Yingfei Liu et al. “PETR: Position embedding transformation for multi-view 3D object detection”. In: (Mar. 2022).

[31] Ze Liu et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”. In: 2021.

[32] Zhijian Liu et al. “BEVFusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation”. In: (May 2022).

[33] Xinzhu Ma et al. “Accurate Monocular Object Detection via Color-Embedded 3D Reconstruction for Autonomous Driving”. In: 2021.

[34] Chen Min et al. “Occ-BEV: Multi-Camera Unified Pretraining via 3D Scene Reconstruction”. In: 2023.

[35] Dennis Park et al. “Is Pseudo-Lidar needed for Monocular 3D Object detection?” In: 2021.

[36] Lang Peng et al. “BEVSegFormer: Bird’s Eye View Semantic Segmentation From Arbitrary Camera Rigs”. In: 2022.

[37] Jonah Philion and Sanja Fidler. “Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D”. In: (Aug. 2020).

[38] Charles R Qi et al. “Frustum PointNets for 3D object detection from RGB-D data”. In: (Nov. 2017).

[39] Charles R Qi et al. “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”. In: (Dec. 2016).

[40] Charles R Qi et al. “PointNet++: Deep hierarchical feature learning on point sets in a metric space”. In: (June 2017).

[41] Cody Reading et al. “Categorical Depth Distribution Network for Monocular 3D Object Detection”. In: 2021.

[42] Thomas Roddick and Roberto Cipolla. “Predicting semantic map representations from images using pyramid occupancy networks”. In: (Mar. 2020).

[43] Thomas Roddick, Alex Kendall, and Roberto Cipolla. “Orthographic Feature Transform for Monocular 3D Object Detection”. In: 2018.

[44] Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. “LMSCNet: Lightweight Multiscale 3D Semantic Completion”. In: (Aug. 2020).

[45] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. “FCAF3D: Fully Convolutional Anchor-Free 3D Object Detection”. In: (Dec. 2021).

[46] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. “ImVoxelNet: Image to Voxels Projection for Monocular and Multi-View General-Purpose 3D Object Detection”. In: 2021.

[47] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. “PointRCNN: 3D object proposal generation and detection from point cloud”. In: (Dec. 2018).

[48] Andrea Simonelli et al. “Disentangling monocular 3D object detection”. In: (May 2019).

[49] Shuran Song et al. “Semantic Scene Completion from a Single Depth Image”. In: 2016.

[50] Pei Sun et al. “Scalability in Perception for Autonomous Driving: Waymo Open Dataset”. In: (Dec. 2019).

[51] Haotian Tang et al. “Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution”. In: 2020.

[52] Xiaoyu Tian et al. “Occ3D: A Large-Scale 3D Occupancy Prediction Benchmark for Autonomous Driving”. In: 2023.

[53] Ashish Vaswani et al. “Attention Is All You Need”. In: 2017.

[54] Tai Wang et al. “FCOS3D: Fully convolutional one-stage monocular 3D object detection”. In: (Apr. 2021).

[55] Tai Wang et al. “Probabilistic and geometric depth: Detecting objects in perspective”. In: *Conference on Robot Learning*. PMLR. 2022, pp. 1475–1485.

[56] Yan Wang et al. “Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving”. In: 2020.

[57] Yue Wang and Justin Solomon. “Object DGCNN: 3D object detection using dynamic graphs”. In: (Oct. 2021).

[58] Yue Wang et al. “DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries”. In: (Oct. 2021).

[59] Yi Wei et al. “SurroundDepth: Entangling Surrounding Views for Self-Supervised Multi-Camera Depth Estimation”. In: 2022.

[60] Yi Wei et al. “SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving”. In: 2023.

[61] Xu Yan et al. “Sparse Single Sweep LiDAR Point Cloud Segmentation via Learning Contextual Shape Priors from Scene Completion”. In: 2020.

[62] Yan Yan, Yuxing Mao, and Bo Li. “SECOND: Sparsely embedded convolutional detection”. en. In: *Sensors (Basel)* 18.10 (Oct. 2018), p. 3337.

[63] Dongqiangzi Ye et al. “LidarMultiNet: Towards a Unified Multi-Task Network for LiDAR Perception”. In: 2023.

[64] Maosheng Ye et al. “DRINet++: Efficient Voxel-as-point Point Cloud Segmentation”. In: 2021.

[65] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. “Center-based 3D object detection and tracking”. In: (June 2020).

[66] Yunpeng Zhang, Zheng Zhu, and Dalong Du. “OccFormer: Dual-path Transformer for Vision-based 3D Semantic Occupancy Prediction”. In: 2023.

[67] Yunpeng Zhang et al. “BEVerse: Unified Perception and Prediction in Birds-Eye-View for Vision-Centric Autonomous Driving”. In: 2022.

[68] Sixiao Zheng et al. “Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers”. In: 2021.

[69] Brady Zhou and Philipp Krähenbühl. “Cross-view Transformers for real-time Map-view Semantic Segmentation”. In: (May 2022).

[70] Yin Zhou and Oncel Tuzel. “Voxelnet: End-to-end learning for point cloud based 3d object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4490–4499.

[71] Xinge Zhu et al. “Cylindrical and asymmetrical 3D convolution networks for LiDAR-based perception”. In: (Sept. 2021).

[72] Xizhou Zhu et al. “Deformable DETR: Deformable Transformers for End-to-End Object Detection”. In: 2021.