Quantify Transformer for End-to-End BEV Scene Understanding

Jia-Yi Zhao [1] , Chui-Hong Chiu [1] , Yu-Chen Lin [1*] , Kuan-Chu Hou [2]
[1] Department of Automatic Control Engineering, Feng Chia University, Taichung 40724, Taiwan, R.O.C
[2] IoT Solution Development at AdasEco Inc.
[*] E-mail(s): yuchlin@fcu.edu.tw (Y. C. Lin).

***Abstract*—Accurate 3D perception is crucial for autonomous driving systems. In this paper, a novel lightweight bird's-eye view (BEV) architecture is introduced, which fully utilizes the camera data and aims to improve the efficiency of perspective conversion and feature representation. The proposed method use a Transformer-like perspective transformation module to efficiently convert camera images into a unified BEV representation, which enables 3D perception including detection, semantic maps in dynamic scenes, and trajectory prediction of surrounding vehicles without relying on LiDAR or other sensors. Experimental results of our proposed model on the nuScenes dataset show improved accuracy and efficiency in integrated 3D perception tasks using camera-only inputs. The ablation study further emphasizes the scalability and adaptability of the framework for real-time 3D perception and environment understanding in complex driving situations.**

Keywords—Autonomous driving, bird's-eye view (BEV), autonomous driving, Transformer, semantic maps

I. Introduction

Automatic driving technology has undergone rapid development in recent years, however, they still have not achieved an accurate enough understanding of the surrounding environment for the safe and reliable operation of self-driving vehicles. The whole problem can be divided into four main modules: perception, prediction, planning and control. The task of perception is to understand the surrounding environment comprehensively, including the recognition and analysis of dynamic objects and static streets, such as parking garages. Surrounding perception capability is the cornerstone of highly automated driving, and is of great significance in improving driving safety and efficiency. Conventional automated driving systems usually rely on a variety of sensors. Due to the advantages of low cost and high performance, pure image-based surrounding perception systems have gradually become the focus of research. Despite the significant progress in 2D and 3D sensing, there are still great challenges in accurately generating bird's-eye view semantic maps in dynamic scenes, performing 3D object detection, and predicting the future behavior of vulnerable road users and vehicles. The complexity of traffic scenes, the interaction of multiple objects, and the variability of environmental conditions, the lack of depth information in image-only sensing systems, further increase the difficulty of accurately understanding 3D space.

To address these challenges, researchers have gradually shifted their research focus to bird's-eye view (BEV) scene understanding, as BEV representations are able to fuse multimodal sensory data in a unified way to visualize spatial layouts and object geometries. BEV methods that utilize only camera data are capable of inferring the semantic information and 3D structure of the surrounding environment, without the need for additional sensor inputs, such as radar or lidar. Unlike traditional methods, image-only BEV systems need to solve the problem of accurate mapping from 2D pixels to 3D scenes. This process involves efficient extraction of pixel-to-space features, aggregation of multi-scale information, and modeling of global spatial context. In particular, existing Convolutional Neural Network (CNN)-based solutions exhibit limitations in dealing with distant objects and complex scenes, which further motivates the application and development of Transformer-based methods in BEV scene understanding.

The focus of this paper is to develop an end-to-end model based on a small Transformer that integrates a 3D object detection module, a semantic graph module, and a future trajectory prediction module. Through a lightweight self-attention mechanism, our model effectively reduces computational cost while maintaining the ability to model global context and detailed features. Together, these modules achieve a comprehensive understanding of the BEV landscape and provide critical support for autonomous driving tasks. In particular, our approach demonstrates significant advantages in data fusion and multi-task learning, which further enhances the model performance and shows some generalization ability to handle complex traffic scenarios.

II. Related Work

A. Camera-Based 3D Object Detection

3D object detection is a crucial component of autonomous driving systems, as it enables vehicles to accurately perceive the location and shape of objects in their surroundings, allowing them to make informed decisions. Traditionally, autonomous driving systems have relied on lidar technology, which provides precise depth information and is particularly effective for object recognition in complex and dynamic environments. However, due to the high cost of lidar and its performance limitations in adverse weather conditions, researchers have increasingly turned to cameras for object detection.

To address these challenges, researchers have proposed RGB image-based 3D detection methods, such as LSS (Lift-Splat-Shoot) [1] and BEVDet [2], which utilize depth estimation to further transform 2D pixel points into 3D spatial features. With the emergence of the Transformer architecture, methods like DETR3D [3] and M2BEV [4] have leveraged the self-attention mechanism to enhance detection accuracy. DETR3D fuses multi-view images to capture global relationships, effectively handling issues like occlusion and changes in illumination, while M2BEV integrates multi-view features into BEV representations to better capture spatial relationships between objects. Additionally, BEVStereo [5] which incorporates temporal information, introduces multi-frame data fusion to infer the motion states and positions of objects through temporal correlations, thereby improving detection performance in dynamic scenes.

B. *Semantic Map*

Bird's-eye view plays a pivotal role in autonomous driving technology by compressing a three-dimensional environment into a two-dimensional representation. This concise, global perspective facilitates vehicle planning, obstacle avoidance, and decision-making. Traditional methods, such as Inverse Perspective Mapping (IPM), convert views into a bird's-eye perspective using a single-stress matrix derived from multi-camera calibrations. However, IPM is prone to distortions caused by non-planar objects or geometrical transformations. PanopticBEV [6] improve upon IPM by addressing these distortions through uni-responsive projections and effectively separating flat and vertical features.

Despite these advancements, single-camera methods face challenges, including feature loss due to low illumination, occlusions, or adverse weather conditions. To overcome these limitations, multi-sensor fusion approach. Since radar systems are expensive, Pseudo-LiDAR offers a cost-effective alternative by estimating depth from cameras and generating radar-like 3D point clouds. Depth estimation methods predict the depth probability of each pixel and convert it into the BEV coordinate system to create top-down viewpoint representations. BEVDepth [7] further optimizes depth estimation accuracy, leveraging multi view fusion aided by neighboring depth information to improve performance in both static and dynamic scenes. Similarly, BEVDet enhances cross-view fusion for efficient BEV representation generation, making it suitable for real-time autonomous driving applications. BEVDet4D [8] extends this by incorporating the time dimension, combining multi-frame information to better capture dynamic object changes and improve detection accuracy.

Transformers have become a popular choice for processing multi-camera images to generate consistent BEV semantic maps by fusing information from different viewpoints. For instance, BEVFormer [9] employs the Transformer attention mechanism to effectively learn spatial and temporal relationships.

C. *Motion Prediction*

Path prediction for dynamic objects is a core challenge in autonomous driving systems. Its goal is to predict the future trajectories of dynamic objects (e.g., vehicles, pedestrians) based on current observation data, providing reliable support for decision-making and path planning. With advancements in sensing technology and computational power, path prediction methods are increasingly adopting deep learning, particularly in the fusion of multimodal data and spatial-temporal feature modeling, showing significant benefit.

Among these, BEV-based path prediction methods have garnered considerable attention due to their ability to more accurately capture the behaviors of dynamic objects and environmental information with a global perspective. FIERY [10], an innovative BEV-based method, leverages exemplary segmentation results and multi-source sensing data for future path prediction. It transforms images into 3D features, projects them onto the BEV plane to align multi-view features and captures dynamic changes. FIERY employs a 3D convolutional temporal module to learn spatial-temporal features and uses a network of convolutional gated recurrent units (Conv-GRU) to predict future multi-view states. Similarly, TBP-Former [11] introduces a Spatial-Temporal Pyramid Transformer (STPT) to accurately predict future BEV states through multi-scale spatial-temporal feature learning and query generation.

To address the challenges of multisensory fusion, long-term prediction, and environmental uncertainty, some approaches explore multi-task joint inference. For instance, BEVerse [12] generates and temporally aligns 4D BEV representations through shared feature extraction and performs joint inference with a multi-task decoder. Unlike FIERY, BEVerse directly predicts potential maps, reducing the impact of uncertainty from different objects and simplifying the learning process to enhance prediction efficiency and accuracy.

In summary, this paper focuses on lightweight BEV landscape understanding that integrates camera-based 3D object detection with semantic mapping information for comprehensive landscape understanding. It can predict the future trajectory of surrounding vehicles, enhance situational awareness and improve driving safety in dynamic driving environments.

III. Methodolgy

This study is divided into three main components. First, the input image is processed by the image backbone to extract essential features and provide key information for subsequent computations. Next, the object detection head is responsible for identifying and localizing objects within the image. Subsequently, the image segmentation module (map head) performs pixel-level classification to accurately distinguish the surrounding environment. Finally, the trajectory prediction module analyzes the motion trajectories of surrounding vehicles and forecasts their future behavior. Fig. 1 illustrates the overall architecture.

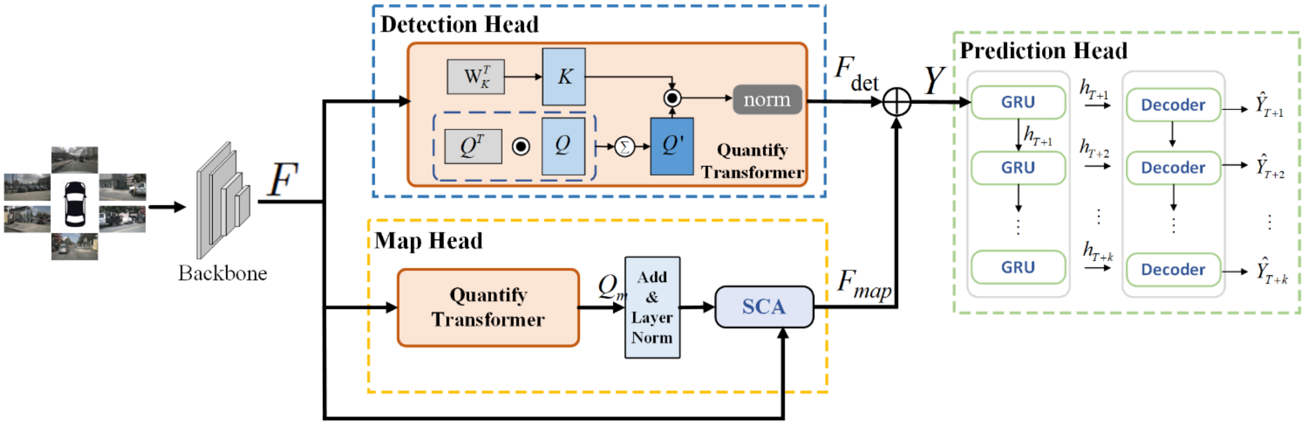

Fig. 1. Overall Architecture.

A. Image Backbone

In this study, ResNet-50 [13] is used as the image backbone network, integrated with a feature pyramid network (FPN) [14] to generate multi-scale feature maps. These maps enable the detection of objects of various sizes and the transformation of these features into BEV representations. The backbone network is initialized by ImageNet [15] pre-training weights to enhance generalization and speed up training. The extracted features are then used as inputs to the converter module to provide key information for spatial-temporal information integration.

B. Quantify Transformer

As shown in Fig. 2, the Quantify Transformer is used as a lightweight and effective enhancement of the traditional self-attention mechanism.

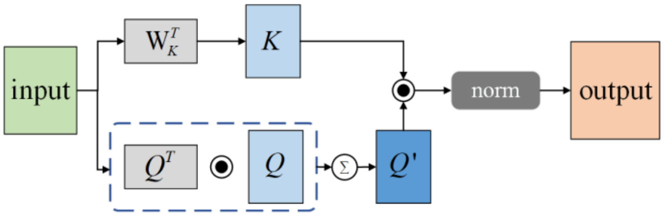

Fig. 2. Quantify Transformer Architecture.

Unlike conventional Transformers, which primarily rely on self-attention to establish intra-sequence relationships, Quantify Transformer introduces an additional self-correlation computation on query, enabling the model to better capture the internal dependencies within query itself. Specifically, in standard Transformers, attention weights are computed based on QK^T and are applied to the V (Value matrix) to retain original features, where $soft\max(\mathrm{QK}^T)$ determines how to weight V. However, in Quantify Transformer, query undergoes a self-correlation computation first, formulated as:

$$Q' = \sum Q * Q^T \tag{1}$$

This operation allows different positions within Q to establish internal connections before interacting with K, thereby enhancing the model's ability to capture semantic information and feature relationships. Consequently, before attending to K, Q already gains a deeper understanding of its own structural dependencies. Furthermore, Quantify Transformer incorporates residual connections to preserve the integrity of the original query matrix Q, mitigating gradient vanishing issues. The complete attention can be expressed as:

$$quantify\ att(Q,K) = norm(Q + K * Q') \tag{2}$$

Unlike conventional Transformers, which compute attention directly on the query and key, the Quantify Transformer first applies self-correlation to Q before interacting with K This pre-processing step enables the query vectors to capture their own structural dependencies, enhancing their expressive power. By refining Q beforehand, the model reduces the search space and computational complexity, leading to a more efficient and lightweight attention mechanism, particularly for high-dimensional inputs.

C. Detection Head

In 3D object detection, we employ the Quantify Transformer to enhance this process, enabling a more structured and computationally efficient attention mechanism. Through the self-correlation of the Quantify Transformer, query representations are refined, allowing the model to learn intra-query dependencies before processing cross-view information. This design is particularly advantageous for multi-view feature aggregation, where image features from different camera angles must be effectively fused to achieve holistic 3D scene understanding. By pre-learning the internal structure of Q, the model enhances spatial awareness capabilities, ensuring robust cross-view feature fusion and more precise object localization even in complex urban environments with occlusions.

For implementation, through the backbone, we map extracted features from multi-view images to the 3D space of each camera. These features serve as queries and keys in the computation. To unify information from different viewpoints, we transform image features to the BEV space via learned geometric transformations, enabling features from different cameras to align under the BEV coordinate system. When performing feature fusion in the BEV space, the Quantify Transformer has already conducted self-correlation learning on Q, allowing information from different viewpoints to correspond more consistently to the BEV feature map, reducing feature misalignment issues. Simultaneously, the query features, having undergone internal structural reinforcement, more effectively capture object boundaries and shape information, improving the accuracy of 3D detection.

D. Map Head

The bird's-eye view semantic map generation process in this thesis takes advantage of the adaptive nature of the Transformer architecture to care for multi-scale features. Specifically, the multiscale features after the encoding-decoding process are first predefined as initial queries that correspond to the size of the BEV plane, where *W* and *H* denote the width and height of the BEV plane, respectively. Next, only the encoded features will be used as inputs for Key and Value, which contain the global information of the scene.

To further improve the accuracy of the feature representation, we introduced a method based on Deformable Attention [16] and Spatial Cross-Attention proposed in BEVFormer. Specifically, the method lifts each query point in the BEV plane to the 3D space, forming query points with 3D position information. These query points are projected back to the 2D view for processing. Due to the limitation of camera parameters, each BEV query can only find valid projection points in 1 or 2 viewpoints, and we use these projection points as reference points to locally sample features around them, and then the BEV query will update these sampled features with weighting, thus realizing spatial feature aggregation. The overall process of Spatial Cross Attention can be represented as follows:

$$\mathrm{SCA}(Q_M, F) = \frac{1}{\left|V_{hit}\right|} \sum_{i \in V_{hit}} \sum_{j=1}^{N_{ref}} \mathrm{DA}(Q_M, P(p,i,j), F^i) \qquad (3)$$

where DA is a deformable attention, i indexes the camera view, j indexes the reference points, and N_{ref} is the total reference points for each query. F^i is the features of the $i\text{-}th$ camera view. For each query Q_M, we use a projection function $P(p,i,j)$ to obtain the $j\text{-}th$ reference point on the $i\text{-}th$ view image. This method can not only handle feature fusion from multiple perspectives but also improve the semantic understanding ability of the scene, thereby generating more accurate bird's-eye semantic maps.

E. Prediction Head

Our prediction head builds upon the features extracted from previous modules, denoted as $Y = \{y_{T-k}, y_{T-k+1}, ..., y_T\}$ where k represents the observation duration. Each y represents the fused features from both the detection head and the map head, which can be expressed as:

$$y = \mathrm{concat}(F_{\det}, F_{map}) \qquad (4)$$

We implement a cascaded structure consisting of multiple Gated Recurrent Units (GRU) coupled with corresponding decoders. GRU modules are arranged in a sequential pipeline, where each unit processes temporal information and passes its hidden state to the next GRU and its paired decoder. Each GRU layer processes the sequence according to:

$$h_{T+1} = \mathrm{GRU}(y_T, h_T) \qquad (5)$$

where h_T is the hidden state from the time step T, and h_{T+1} represents the hidden state of the GRU layer at the next time step. Then, each GRU-decoder processes the hidden state and maps it into the prediction space to produce the output $\hat{Y}$ for the $l\text{-}th$ future time step:

$$\hat{Y}_{T+l} = \mathrm{Decoder}(h_T, l) \qquad (6)$$

The prediction head operates by leveraging GRU-decoders to extract temporal relationships from the hidden states. For each future time step, it generates outputs positions through simple linear layers. This sequential design allows the model to effectively capture and utilize temporal dependencies, producing accurate trajectory forecasts while maintaining computational efficiency.

IV. Experiments

A. Dataset and Evalution Metrics

In this paper, the nuScenes [17] dataset for experiments and model validation is used. The nuScenes is an open-source dataset widely used in the field of automated driving research, providing rich multimodal data. For the fairness of the experiments, our experimental results are evaluated and compared on the validation set. To fully evaluate our method, we used the official evaluation metrics provided by nuScenes, which can fully reflect the performance of the model in various aspects. For 3D detection tasks, we report the nuScenes Detection Score (NDS), mean Average Precision (mAP), and five True-Positive metrics, including Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE) and Average Attribute Error (AAE). The semantic categories constructed in the semantic maps included lane dividers, pedestrian crossings and lane boundaries. For quantitative evaluation, we calculated the mean intersection over Union (mIOU) for each category between the prediction and the ground truth map as a ranking indicator. In the field of trajectory prediction, we followed the rules set by nuScenes for evaluation, including minimum average displacement error (ADE), minimum final displacement error (FDE), and loss rate greater than 2 meters, and used information from the last 2 seconds to predict future trajectories.

B. Results

We evaluate our model on the nuScenes validation set and compare its performance with existing methods. As shown in Table I, our model, using only camera images as input, achieves 58.6 mAP and 66.0 NDS in 3D object detection. Table II reports 49.8 mIOU in map semantic segmentation, surpassing previous approaches and establishing a new benchmark for this task. Additionally, as illustrated in Fig. 2, the visualization results demonstrate the model's perception and prediction capabilities

TABLE I. 3D DETECTION RESULTS ON NUSCENES VAL SET.

Method	Modality	Backbone	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
DETR3D (2021)	C	R101	0.346	0.425	0.811	0.282	0.493	0.979	0.212
BEVerse-Tiny (2022)	C	Swin-T [18]	0.321	0.466	0.681	0.278	0.466	0.328	0.190
BEVDet4D (2022)	C	R101	0.421	0.545	0.579	0.258	0.329	0.301	0.191
BEVFormer (2022)	C	R101	0.416	0.517	0.673	0.274	0.372	0.394	0.198
TSC-BEV (2023) [19]	C	R50	0.399	0.521	0.557	0.272	0.464	0.287	0.205
RCBEV4d (2023) [20]	C+R	Swin-T	0.381	0.497	0.526	0.272	0.445	0.465	0.185
SparseDrive (2024) [21]	C	R101	0.496	0.588	0.543	0.269	0.376	0.229	**0.179**
HENet (2024) [22]	C	V2-99 [23] & R50	0.502	0.599	0.465	0.261	0.335	0.267	0.197
Ours	C	R50	**0.586**	**0.660**	**0.335**	**0.263**	**0.289**	**0.248**	0.190

recognition, further validating its accuracy and applicability. These results indicate that our model remains competitive even without leveraging depth information.

While a detailed comparison of the motion prediction results shown in Table III, which indicates that our model achieves the lowest error rate, demonstrating its strength in minimizing false predictions, we still have room for further improvement in this area. Nevertheless, the overall results highlight the versatility and robustness of our approach, especially given its lightweight design and resource efficiency.

TABLE II. MAP SEGMENTATION ON NUSCENES VAL SET.

Method	Divider	Ped Cross	Boundary	mIOU↑
BEVFormer (2022)	42.1	23.8	41.6	35.9
BEVerse (2022)	53.2	39.0	53.9	48.7
Bi-Mapper (2023) [24]	43.8	25.7	44.2	37.9
DriveWorld (2024)	29.5	17.2	34.2	27.0
Ours	**53.7**	**39.8**	**55.9**	**49.8**

C. Ablation Studies

The ablation studies in Table IV confirm that that Quantify Transformer not only contributes significantly to model performance but also greatly enhances the model lightweight characteristics. Especially in the detection and map parts. In our experiments, we replaced a standard Transformer with the Quantify Transformer while conducting tests on an NVIDIA RTX 4090 GPU. The results demonstrate notable improvements across multiple metrics: the model parameter count decreased from 226.28M to 181.18M, representing a 20% reduction in model size. Simultaneously, the computational requirements were reduced from 168.95 GFLOPS to 112.85 GFLOPS, indicating a 33% decrease in computational complexity. These optimizations led to improved inference efficiency, with the model processing speed increasing from 5.8 FPS to 7.3 FPS, marking a 26% enhancement in throughput. These comprehensive improvements highlight the effectiveness of the Quantify Transformer in achieving both model compression and inference acceleration while maintaining performance.

TABLE III. MOTION PREDICTION ON NUSCENES VAL SET.

Method	minADE(m)↓	minFDE(m)↓	MR↓	EPA↑
UniAD (2023)[25]	0.71	1.02	0.151	0.456
DriveWorld (2024)	0.61	**0.91**	0.136	0.503
SparseDrive (2024)	**0.60**	0.96	0.132	**0.555**
Ours	0.83	1.14	**0.128**	0.525

TABLE IV. QUANTIFY TRANSFORMER PERFORMANCE COMPARISON.

Method	mAP↑	NDS↑	Params(M)	GFLOPs	FPS
Ours (w/o)	0.461	0.404	226.28	168.95	4.7
Ours (w/)	0.586	0.660	181.18	112.85	7.3

V. CONCLUSION

This paper proposes an end-to-end BEV scene understanding model based on Quantify Transformer and successfully integrates 3D object detection, speech intent generation and future trajectory prediction modules. By introducing a lightweight self-attention mechanism, our model significantly reduces the computational cost while maintaining the ability to accurately model complex scenes and detailed features. This enables the model to run efficiently and accurately in real-world applications such as autonomous driving. Experimental results show that our approach improves several key performance metrics compared to baseline methods, especially in terms of accuracy, inference efficiency, and computational cost, and achieves higher operational efficiency.

ACKNOWLEDGMENT

This Research was supported by National Science and Technology Council, R.O.C., under the Grant NSTC 112-2628-E-035-001-MY3.

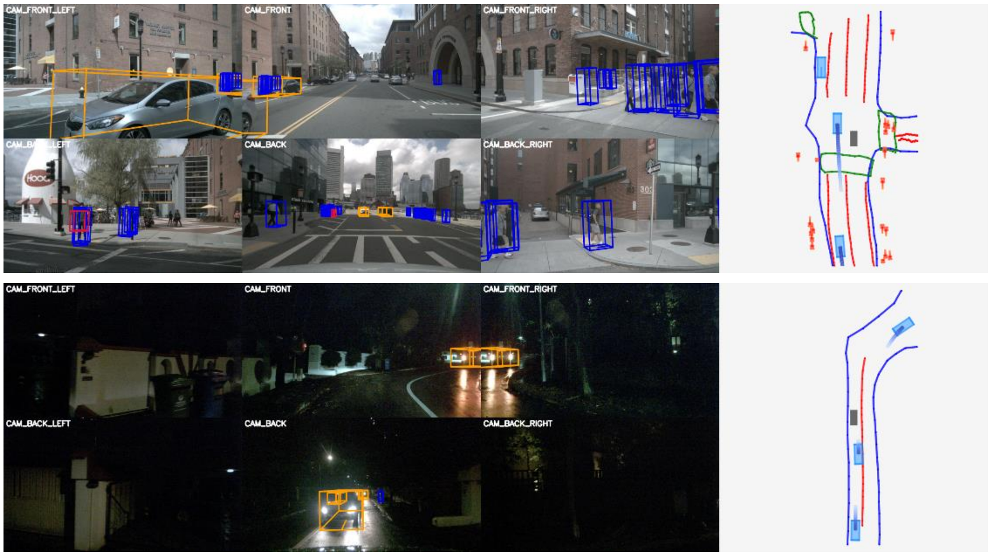

Fig. 3. Visualization Result.

References

[1] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," *European Conference on Computer Vision (ECCV)*, pp. 194-210, 2020.

[2] J. Huang, G. Huang and Z. Zhu *et al*., "BEVDet: High-performance multicamera 3D object detection in bird-eye-view," *arXiv:2112.11790*, pp. 1-19, 2022.

[3] Y. Wang, V. Guizilini and T. Zhang *et al*., "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," *arXiv:2110.06922v1*, pp. 1-12, 2021.

[4] E. Xie, Z. Yu and D. Zhou *et al*., "M2BEV: Multi-camera joint 3D detection and segmentation with unified bird's eye view representation," *arXiv:2204.05088*, pp 1-21, 2022.

[5] Y. Li, H. Bao and G. Zheng *et al*., "BEVStereo: Enhancing depth estimation in multi-view 3D object detection with temporal stereo," *AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp.1486-1494, 2023.

[6] N. Gosala and A. Valada, "Bird's-eye-view panoptic segmentation using monocular frontal view images," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1968-1975, 2022.

[7] Y. Li, G. Zheng and G. Yu *et al*., "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," *AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 1477-1485, 2023.

[8] J. Huang and G. Huang, "BEVDet4D: Exploit temporal cues in multi-camera 3D object detection," *arXiv:2203.17054*, pp. 1-11, 2022.

[9] Z. Li, W. Wang and H. Li *et al*., "Bevformer: Learning bird's-eye-view representation from multicamera images via spatiotemporal transformers," *European Conference on Computer Vision (ECCV)*, pp. 1-18, 2022.

[10] A. Hu, Z. Murez and N. Mohan *et al*., "FIERY: Future instance prediction in bird's-eye view from surround monocular cameras," *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, pp. 15253-15262, 2021.

[11] S. Fang, Z. Wang and Y. Zhong *et al*., "TBP-Former: Learning temporal bird's-eye-view pyramid for joint perception and prediction in vision-centric autonomous driving." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, pp. 1368-1378, 2023.

[12] Y. Zhang, Z. Zhu and W. Zheng *et al*., "BEVerse: Unified perception and prediction in bird's-eye-view for vision-centric autonomous driving." *arXiv:2205.09743*, pp.1-12, 2022.

[13] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 770-778, 2016.

[14] T.Y. Lin, P. Dollár and R. Girshick *et al.*, "Feature pyramid networks for object detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, pp. 936-944, 2017.

[15] A. Krizhevsky, I. Sutskever and G. E Hinton, "Imagenet classification with deep convolutional neural net works," *Neural Information Processing Systems*, pp. 1097-1105, 2012.

[16] X. Zhu, W. Su and L. Lu *et al.*, "Deformable detr : Deformabletransformers for end-to-end object detection," *International Conference on Learning Representations*, pp. 1-16, 2021.

[17] H. Caesar, V. Bankiti and A. H. Lang *et al.*, "nuScenes: A multimodal dataset for autonomous driving," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, pp. 11621-11631, 2020.

[18] Z. Liu, Y. Lin and Y. Cao *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, pp. 9992-10002, 2021.

[19] J. Fang, M. Zhuang and G. Wang *et al.*, "TSC-BEV: Temporal-Spatial feature consistency 3D object detection," *China Automation Congress (CAC)*, Chongqing, China, pp. 6899-6904, 2023.

[20] T. Zhou, J. Chen and Y. Shi *et al.*, "Bridging the view disparity between radar and camera features for multi-modal fusion 3D object detection," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1523-1535, 2023.

[21] W. Sun, X. Lin and Y. Shi *et al.*, "SparseDrive: End-to-End autonomous driving via sparse scene representation," *arXiv: 2405.19620*, pp. 1-16, 2024.

[22] Z. Xia, Z. Lin and X. Wang *et al.*, "HENet: Hybrid encoding for end-to-end multi-task 3D perception from multi-view cameras," *European Conference on Computer Vision (ECCV)*, pp. 376-392, 2024.

[23] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, pp. 13903-13912, 2020.

[24] S. Li, K. Yang and H. Shi *et al.*, "Bi-Mapper: Holistic BEV semantic mapping for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 7034-7041, 2023.

[25] Y. Hu, J. Yang and L. Chen *et al.*, "Planning-oriented autonomous driving," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, pp. 17853-17862, 2023.