BEV-Based 3D Detection for Automatic Driving Using Lidar-Camera Fusion

Jihua Jiang
College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, 211800, China

Keywords: Bird's Eye View, 3D Detection, Lidar-Camera Fusion, Automatic Driving.

Abstract: At present, deep learning technology and automatic driving related research are gradually mature. Autonomous driving perception technology has been developed tremendously as an important part of the autonomous driving system. This paper explores the development of a 3D detection task based on the Lidar-Camera fusion (LC Fusion) scheme of BEV technology from different fusion mechanisms. This paper concludes that the LC Fusion algorithm for BEV will be the most promising perception approach at present and the main form of perception system in the future. The BEV-based LC Fusion has many advantages such as high detection accuracy and robustness. Starting from the fusion granularity, this paper summarizes the characteristics of high accuracy and low latency of the current LC Fusion algorithm and the limitations such as network complexity, and proposes improvements such as attention fusion mechanism and network lightweighting for the problems faced by the algorithm. In addition, this paper proposes solutions such as unified spatial representation and decoupling of sensing channels, as well as the development direction of sensing systems including end-to-end design, multi-task learning, and knowledge distillation. This paper can provide reference materials and summarize perspectives for subsequent related researchers to pave the way for the development of this perception technology.

1 INTRODUCTION

In recent years, neural networks and deep learning technologies have developed rapidly. In addition, automatic driving perception technology, as the foundation of the whole automatic driving system, plays a key role in the subsequent decision-making and control. Due to the excellent detection accuracy and rich scalability, multi-sensor fusion perception schemes have gradually become mainstream, among which the Lidar-Camera Fusion scheme has been widely noticed by academia and industry for its highly complementary characteristics and excellent perception performance.

At this stage, Lidar-Camera Fusion (LC Fusion) can be categorized into pre-fusion, cascade fusion, and post-fusion according to the fusion stages. Each of these three fusion stages possesses its characteristics and is a hot research topic at present.

Bird-eye-view (BEV) as a more popular perception method has received widespread attention. BEV perception technology is to convert sensory information into features in BEV space, and due to the uniformity of the BEV space, it is easier to realize the fusion and processing of multiple heterogeneous modal data as well as multitasking learning and other tasks. Not only that, the overlapping of objects can be reduced in the BEV perspective to get a clearer and more accurate perceptual environment. In addition, BEV perception can also be directly optimized end-to-end by neural network algorithms, without the need for serial perception channels, which can avoid the accumulation of errors, as well as reduce the impact of the algorithmic logic and improve the accuracy of perception and prediction.BEV perception provides a strong perceptual foundation for important functions such as decision-making and control, which is of great significance for the eventual realization of high-level automated driving.

Therefore the addition of BEV spatial technology provides new solution ideas for LC Fusion. Unlike traditional perception schemes, BEV perception schemes do not stack sub-task modules in a linear structure but convert camera and radar sensing to a unified bird's-eye view for related perception tasks, which will give BEV perception the advantages of easier cross-camera and multi-modal fusion, as well as the realization of timing fusion.

Jiang, J.
BEV-Based 3D Detection for Automatic Driving Using Lidar-Camera Fusion.
DOI: 10.5220/0012838400004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 552-559
ISBN: 978-989-758-690-3

This paper mainly focuses on the technical characteristics and problems faced by the existing Lidar-Camera fusion schemes based on BEV technology, to obtain the focus and development direction of the existing fusion technology, to provide a summarized perspective for researchers in this field, and to provide research ideas and focus for later researchers. This paper introduces the concept of BEV technology as well as its advantages and summarizes the perception methods based on BEV according to the different sensors, in which the advantages and disadvantages of the sensors are also compared and analyzed. Then, this paper provides a systematic review of BEV-based Lidar-Camera fusion algorithms, details the advantages and disadvantages of the relevant representative algorithms at three fusion granularities: point-level fusion, feature-level fusion, and voxel-level fusion, and summarizes the common challenges and solutions faced by multimodal fusion algorithms. Finally, this paper discusses the future research direction of multimodal fusion perception methods.

2 THE CLASSIFICATION OF BEV

Depending on the input modality, BEV perception algorithms can be categorized into pure LiDAR-based, pure vision-based, and multimodal fusion-based approaches. This section focuses on these types of perception algorithms.

2.1 BEV Perception Algorithms Based on Pure LiDAR

LiDAR calculates the distance to the surrounding objects by measuring the time difference between the laser beam being emitted and being received by the object, which can realize accurate sensing of geometrical physical information such as object attitude, shape, and speed. This is important for intelligent vehicles to understand the surrounding environment in real time. However, the sparseness of the LiDAR point cloud makes it unable to bring rich semantic information. Currently, Lidar also has limitations such as poor long-distance sensing accuracy and expensive price. In recent years, Lidar-based sensing algorithms have gained significant development, realizing sensing accuracy that cannot be achieved by other sensors.

Pure Lidar algorithms can be categorized into point-based, voxel-based, and feature projection-based methods. Examples include the classical pointnet as well as voxelnet (Qi et al. 2017, Zhou & Tuzel 2018).

2.2 Pure Vision-based BEV Perception Algorithms

Pure visual images obtained from cameras can provide dense information, display shape and texture attributes, and contain a large amount of rich semantic information. However, image data depth information is lost and it is difficult to obtain accurate geometric information such as distance. Although network research for image depth estimation has been developing rapidly recently, it is still difficult to realize accurate depth estimation, which severely limits the detection accuracy. Purely visual BEV perception algorithms can be categorized into MLP-based, Transformer-based, and depth-based methods based on the transformation from perspective view (PV) to bird's eye view (BEV) (Ma et al. 2022). The Transformer-based and depth-based perception algorithms, which have been studied more, are represented by algorithms with superior performance, such as BEVFormer, BEVDet4D, and so on (Li et al. 2022, Huang et al. 2022).

2.3 BEV Perception Algorithm based on LC Fusion

Due to the highly complementary characteristics of LiDAR and camera, combining the two for perception has become a mainstream research direction. LiDAR can obtain physical information such as object contour, distance, etc., as a way to make up for the camera's difficulty in predicting the accurate depth. The RGB map of the camera can provide rich semantic information, LiDAR provides physical information, such as object contours and distances, complementing the camera's difficulty in predicting accurate depth. The RGB map from the camera offers rich semantic information, compensating for LiDAR's lack of semantics and the challenge of recognizing distant objects. This fusion enhances detection accuracy and robustness. Currently, the three fusion methods have been thoroughly studied, and many research teams have proposed models with powerful performance, such as TransFusion and BEVFusion (Bai et al. 2022, Liang et al. 2022).

3 BEV-BASED LC FUSION

The following is a categorization and specific description of BEV-based LC Fusion algorithm. It can be specifically categorized into point-level-based fusion, feature-level fusion, and voxel level fusion.

3.1 Point-Level Fusion

Point-level fusion is the fusion of Lidar points and camera images at the data input stage, which has more complete data preservation and less information loss. At this stage, point-level fusion first finds the association between Lidar points and image pixels based on the calibration matrix and then augments the Lidar features with segmentation scores or CNN features by concatenating the associated pixels point by point. For example, PointAugmenting projects the Lidar points onto the image plane and decorates the Lidar points using semantics in the camera image to make them better recognizable for long-range and occluded targets (Wang et al. 2021). This model is a classic example of point-level fusion, which realizes the alignment of LIDAR points with image data at the semantic level in a real sense, although there is still much room for improvement.

There is also the equally classic PiontPainting, which also uses point-by-point splicing to realize fusion (Vora et al. 2020). It fully utilizes the complementary characteristics of point cloud data and image data, using the image semantic segmentation classification results to be spliced with Lidar points, and decorating Lidar features with the semantic segmentation scores, and finally a Lidar-based object detector can be used on this decorated point cloud to obtain 3D detection.

Despite the improvements made, these point-level fusion methods still suffer from two main problems. First, they only fuse Lidar features and image features by point-by-point summation or concatenation, and therefore and their dependence on image quality. Second, finding hard correlations between sparse Lidar points and dense image pixels not only wastes many image features with rich semantic information, but also relies heavily on high-quality alignment between the two sensors, but the inherent spatio-temporal mismatch makes it more difficult to achieve high-quality alignment.

In addition, in recent years, there have also been ways to convert image data into dense pseudo-point clouds and then fuse them with Lidar point clouds to obtain perceptual results in a Lidar 3D target detection backbone. In 2021, Yin et al. proposed MVP, which introduced the fusion of Lidar data using virtual points (Yin et al. 2021). In 2023, Wu et al. proposed Vir ConvNet, a method that uses the VirConv module for virtual point-based 3D object detection, addressing issues of dense virtual points and noise introduced by depth complementation (Wu et al. 2023). Currently, this fusion method using the conversion of images into pseudo-point clouds has received widespread attention, and it has the advantage of higher accuracy of 2D image depth complementary network and better utility and generalization ability, which provides a new solution idea for the future fusion perception scheme.

3.2 Feature-Level Fusion

Feature-level fusion refers to the fusion of sensory information with feature data of different modalities after obtaining relevant abstract features through neural networks to obtain fused features for subsequent processing and related sensory tasks.

A more well-known feature-level fusion at this stage is TransFusion proposed in 2021, which consists of a convolutional backbone and a detection head based on a Transformer decoder (Bai et al. 2022). This model enables the use of sparse object query sets to predict the initial bounding box from a LiDAR point cloud guided by image features while employing an attentional mechanism to adaptively find important features and fuse them. The TransFusion model achieves soft connectivity at the feature level as opposed to the hard connectivity of the Lidar points projected into the image in the point-level fusion, which effectively prevents image degradation and sensor misalignment due to the quality degradation and sensor misalignment, and has better robustness in terms of perceptual degradation.

In 2022, Chen et al. proposed AtuoAlign (Chen, et al. 2023). Instead of using camera projection matrices and geometric transformations to establish deterministic correspondences, the model uses a learnable alignment map to model the mapping relationship between the image and the point cloud.AtuoAlign achieves positional and semantic consistency at both pixel-level and instance level, which effectively ensures the accuracy of feature alignment at different granularities. However, the cross-attention feature alignment module in the AtuoAlign model inevitably brings the cost of high computation because it adopts the global attention mechanism. Therefore, the AtuoAlign model can hardly bear the computational cost of querying high-resolution images, and can only be limited to lower-resolution and lower-quality images, which reduces its performance.

The team proposed AtuoAlignV2 to solve this problem. This model proposes the cross-modal DeformCAFA model (Chen et al. 2024). It takes into account the sparse learnable sampling points for cross-channel relational modeling, enhances the tolerance of calibration error, accelerates the feature aggregation between different channels, improves computational efficiency, and alleviates the contradiction. Meanwhile, the model proposes the GT-AUG data enhancement model. It can handle the target occlusion problem in the image domain and generate smoother images for fusion simply and efficiently. The team also considered the problem of degradation of perception accuracy due to missing images for image inputs in multi-view perception in real-world scenarios and proposed an image discarding strategy, which will improve the training speed of the model because fewer images are processed in each batch, and will improve the overall performance and robustness of the model. However, network lightweight design is still a key point to consider for its future practical deployment.

3.3 Voxel-Level Fusion

In recent years, voxel-level fusion has been developed more rapidly. The main feature of this method is to decouple the Lidar and Camera perceptual streams, and the two modalities are subject to independent feature extraction and prediction, and finally fused in voxel space. This fusion approach is simple and efficient, reduces the model's dependence on the complete modal input, and improves the robustness of the perceptual system.

In 2022, BEVFusion proposed by Liang et al. drew attention to this fusion method (Liang et al. 2022). This model truly realizes the decoupling of the two perceptual streams and solves the problem that multimodal inputs are highly dependent on the complete inputs.BEVFusion proposes to decouple the camera and the Lidar to form two independent perceptual streams so that their raw data are converted into BEV spatial features through the BEV encoder and finally fused using a simple module. Such a simple design makes it possible to directly use mature models at this stage for relevant perceptual tasks, which makes its generalization ability greatly enhanced.

To achieve a more robust and accurate perception capability, Ge et al. proposed MetaBEV in 2023, an algorithm that possesses strong capabilities in dealing with feature alignment and sensor failure problems (Ge et al. 2023). The model proposes a cross-modal BEV Evolving decoder based on the independent perception module of BEVFusion, which uses cross-modal variable attention to aggregate learnable queries with Camera BEV features and Lidar Bev features to obtain fused features. Finally, several task-specific heads are applied to support different 3D perception tasks. A robust fusion module with a new M2oE-FFN layer is introduced. The main purpose is to mitigate gradient conflicts between 3D target detection and semantic segmentation tasks for more balanced and robust performance.MetaBEV achieves multi-task perception in multimodal fusion. Although MetaBEV improved BEVFusion, it still used two modally different BEV encoders, which failed to completely solve the problem of feature misalignment.

So in 2023 wang et al. proposed UniBEV, which is a model that differs from previous multimodal detection methods based on a unified BEV representation in that all sensor modalities will use a unified BEV encoder based on variable attention to extract BEV features from the original coordinate system and enable the features to be fused in the channel normalized weights module (Wang et al. 2023). This can properly solve the data discrepancies due to inconsistent encoding methods. More importantly, the UniBEV model designs the channel normalized weight module (CNW) to fuse the two perceptual streams. This module fuses the BEV feature maps by summing and averaging all available modal feature maps. The averaging can dilute the information from the more reliable sensors with the information from the less reliable sensors, alleviating the contradiction caused by the different sensor information. Therefore, this fusion mechanism makes the model more robust for dealing with sensor faults or failures.

In addition, in recent years knowledge distillation techniques have received increasing attention for their advantages in the fusion of heterogeneous non-uniform features in both 2D and 3D spaces. In 2023, Chen et al. proposed BEVDISTILL (Chen et al. 2022). The model unifies image and Lidar features in BEV space and enables fusion by adaptively transferring knowledge through dense and sparse feature distillation of non-homogeneous representations in a teacher-student paradigm. Notably, the model's distillation paradigm takes note of the differences between modalities. Knowledge distillation is a different perceptual approach from the conventional use of attentional mechanisms to achieve alignment and fusion between two modalities, and it has advantages for achieving alignment between heterogeneous non-uniform data,

which has become an important development direction for future multimodal fusion.

There are also methods such as UVTR for knowledge distillation fusion at the voxel level, which avoid problems such as semantic ambiguity caused by compression by transferring knowledge between the student model and the teacher model at the voxel level without a high degree of compression (Li et al 2022).

4 ADVISE

The fourth subsection will discuss three problems and related solutions faced by BEV-based multimodal fusion detection algorithms, as well as future directions for centralized development.

4.1 Challenges and Possible Solutions

For models based on BEV space for multimodal fusion detection, three problems will be commonly faced, which are the difficulty of fusion feature alignment, high dependence on the complete modal input, inaccurate depth information, and semantic ambiguity due to depth estimation or spatial compression of image data or point cloud data. The solutions provided by different algorithmic models are discussed below starting from the problems.

4.1.1 Difficulty in Fusion Feature Alignment

Since multi-sensor fusion involves the fusion of sensory information with different extraction methods and different data forms, the problem of difficult feature alignment of different modal data is bound to occur in the fusion process. This can lead to the introduction of errors, resulting in problems such as decreased sensing accuracy.

First, the algorithm represented by DeepFusion proposes two techniques, InverseAug and LearnableAlign, to achieve robust alignment and fusion (Meng et al. 2022). These two techniques bring low computational costs and provide large gains in recognizing long-range targets, significantly improving the model's recognition and localization capabilities.

The second is the algorithm represented by UniBEV. In UniBEV, the research team designed a variable-attention-based unified BEV encoder, which would enable high-precision alignment of heterogeneous data features in the BEV domain. The model improves on BEVFusion as well as MetaBEV by utilizing a variable attention module on Lidar and Camera perceptual streams to achieve a unified encoding of BEV features to facilitate the interaction of the information between the two and finally achieve adaptive fusion of the two through the channel normalized weight module (CNW). It can be seen that the unified representation-based approach has a natural advantage in solving the problem of feature alignment.

The third is the algorithm represented by AutoAlign (Chen, et al. 2023). The AtuoAlign algorithm is designed with a cross-attention feature alignment module (CAFA), and a self-supervised cross-channel feature interaction module (SCFI). With these two modules, heterogeneous data features can be aligned in a dynamic and data-driven manner. It can be shown that the current adaptive feature alignment methods based on attention mechanisms as well as those based on uniform spatial representations are accurate and efficient. In recent years, there is a large potential for research in this direction. However, performing network lightweight as well as low-latency design are still issues that need to be considered in this area.

4.1.2 High Dependence on Complete Modal Inputs

Many current multimodal fusion algorithms rely heavily on the accurate sensing of Lidar points, with the Lidar point cloud detection network as the backbone. This leads to the problem of severe degradation of perception when the Lidar sensor fails or malfunctions, and even threatens the safe driving of self-driving cars. It can be said that this is a problem that needs to be solved urgently at present for the development of autonomous driving. In recent years, many research teams have given suitable solutions to this problem.

The first one is the algorithm represented by BEVFusion.BEVFusion adopts the method of decoupling the Lidar and Camera sensing streams, which can reduce the heavy dependence on Lidar in fusion, as well as effectively prevent the problem of sensor failure. This ensures that when one sensor channel fails, the task can be continued by another perceptual channel. Later on, a research team successively proposed MetaBEV and UniBEV based on the BEVFusion model, which further improved the performance, but this inevitably leads to the problem of perceptual degradation when failure occurs.

Next is the algorithm represented by Policy Fusion (Huang et al. 2023). This algorithm is a decision-level fusion algorithm that utilizes the VCNet network (Value Critique Network) in the

decision-making phase to score the content of the decisions obtained from end-to-end to obtain the primary and secondary decisions. The two decisions are then fused by the Primary and Secondary Strategy Fusion (PSF) module. In case of severe sensor failure, the unaffected independent decision will replace the fused decision. Sensor failures as well as failures can be solved properly.

4.1.3 Inaccurate Depth Information and Semantic Ambiguity Due to Depth Estimation or Spatial Compression of Image Data or Point Cloud Data

When depth estimation is performed on image data from two-dimensional space to three-dimensional space, the problem of inaccurate depth information is caused by the limited prediction accuracy of the depth estimation network. When compression of LiDAR point cloud data into BEV space is performed, the high degree of compression at each location aggregates features from different objects, which can lead to problems such as semantic ambiguity.

To solve this problem, the research team proposed an algorithm represented by UVTR, which uniformly encodes inputs from different sensors and performs a uniform representation of different modal data in voxel space without further compression (Li et al 2022). However, the problem of heterogeneous data characterization is ignored in such methods and the two are fused by forced knowledge distillation. This is the direction for further enhancement of such methods.

4.2 Future Research Direction

This section will discuss several future directions of multimodal fusion perception approaches, including end-to-end, multi-task learning, and temporal fusion.

4.2.1 End-to-End

End-to-end perception is to make the original sensory data through the neural network directly output the prediction results or other task results, no longer needing to pre-process the original data for feature capture. This kind of perception can avoid manually designed features, but use the deep learning neural network to adaptively extract effective features for related perception tasks, which can give the model more space to automatically adjust according to the data and improve the robustness of perception. And the end-to-end design can reduce the manual intervention and post-processing steps, reducing error accumulation.

4.2.2 Multi-Task Learning

Multi-task learning (MTL) refers to the execution of multiple task predictions with a set of trained weights, which receives widespread attention due to its practical value, such as complementary performance and lower computational cost. However, in multitask learning models must learn to balance the various objectives of each task and avoid task conflicts, which will be challenging.MetaBEV proposed a robust fusion module with a new M2oE-FFN layer (Ge et al. 2023). Its main role is to mitigate conflicts between multiple tasks. Experiments show that the method has a better balancing effect. However, further improvements are still needed to realize a more concise and efficient multi-task learning network.

4.2.3 Timing Fusion

Timing fusion is the key to improving the accuracy and continuity of the perception algorithm, improving the problems of inter-frame jumps and target occlusion in target detection, more accurately determining the target motion speed, and also playing an important role in target prediction and tracking. Currently, temporal fusion has become the direction of attention in perception modeling. In 2022 BEVDet4D was proposed by Huang et al (Huang et al. 2022). It makes the features of the current frame and the BEV features of the previous frame first go through the alignment in the temporal and spatial dimensions and then spliced in the channel dimension. This method substantially improves the performance in speed prediction through temporal fusion with only a small increase in computational cost.

4.2.4 Low Latency

Most of the current multimodal fusion algorithms for real commercial scenarios have computational efficiency and computational volume that are difficult to deploy to vehicle ECUs for reliable real-time sensing. Low latency has become an urgent problem to be solved for the practical deployment of sensing models. In 2021, a research team proposed a Lidar-based real-time detection model BEVDetNet (Mohapatra et al. 2021). With a concise and efficient design, the model truly realizes low-latency detection. It has a latency of 4ms on the embedded Nvidia Xille platform, which can reach the level for commercial deployment. To have real-time performance,

reasonable network design and control of the amount of computation are essential. Achieving a balance between detection accuracy and real-time performance is a key consideration for future perceptual models.

4.2.5 Knowledge Distillation

The purpose of knowledge distillation is to migrate the knowledge learned from a large model or multiple model integration to another lightweight model. It is essentially a method of model compression. This method can achieve a significant reduction in model size without reducing the detection accuracy of the original model, which can make the model commercially deployable. Knowledge distillation is the transfer of information between the teacher model and the student model. Specifically, knowledge is suggested from the trained teacher model to the lightweight student model. This approach reduces the model size while keeping the detection effectiveness. Recently, BEVDISTILL (Chen et al. 2022) performed two feature distillations, dense and sparse, between teacher and student models for feature alignment optimization and knowledge migration at the instance prediction level. The model was highly successful on the nuScenes dataset and also proved that knowledge distillation is a powerful tool for solving practical deployment challenges.

5 CONCLUSION

Due to the increasing importance of accurate and robust perception techniques in applications such as autonomous driving, this paper explores multimodal fusion 3D target detection algorithms based on BEV technology, focusing on the fusion of LIDAR and camera vision for perception. First, this paper derives the advantages of BEV technology in current perception systems, including occlusion reduction, end-to-end design, and error accumulation reduction. By analyzing the advantages and disadvantages of BEV sensing algorithms and sensors for image-only, LIDAR-only, and LC fusion, this study finds that the LC fusion sensing method has better detection accuracy and robustness, and is one of the most promising sensing methods in the future. In addition, the advantages and limitations of the related algorithmic models are discussed from the three fusion granularities of point-level fusion, feature-level fusion, and voxel-level fusion, among which the algorithms based on virtual points, the algorithms based on the unified representation of the BEV space, and the algorithms related to the distillation of knowledge are of pioneering significance. Aiming at these models, this paper puts forward suggestions such as network lightweight design. However, the multimodal fusion approach faces three challenges, including difficulties in fusion feature alignment, high dependence on complete modal inputs, and depth estimation or spatial compression leading to inaccurate depth information and semantic ambiguity. Approaches such as unified spatial representation and decoupling of perceptual channels are suggested to address these issues. In the future, multimodal fusion perception methods can evolve towards end-to-end, multi-task learning, temporal fusion, low latency, and knowledge distillation. This paper is dedicated to exploring the characteristics of multimodal fusion techniques and mining the potential development directions, which provides a reference and summary perspective for future research.

REFERENCES

C. R.Qi, H., K. C.Mo, et al, *PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation*; in proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, F Jul 21-26, (2017).

C. W.Wang, C., M.Zhu, et al, *PointAugmenting: Cross-Modal Augmentation for 3D Object Detection*; in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, F Jun 19-25, 2021. 2021.

C.Ge, J.Chen, E,Xie, et al, ArXiv.2304.09801, (2023).

H.Wu, C. L.Wen, S S.Shi, et al, *Virtual Sparse Convolution for Multimodal 3D Object Detection*; in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, CANADA, F Jun 17-24, (2023).

J. J. Huang, et al, *BEVDet4D: Exploit Temporal Cues in Multi-camera 3D Object Detection*; arXiv abs/2203.17054(2022)

S. Mohapatra, S.Yogamani, H.Gotzig, et al, 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), 2809-15(2021).

S.Vora, A H.Lang, B.Helou, et al, *PointPainting: Sequential Fusion for 3D Object Detection*; in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, F Jun 14-19, 2020. 2020.

S.Wang, H. Caesar, L Nan, et al, ArXiv.2309.14516, (2023).

T. W.Yin, X Y.Zhou, P KRäHENBHüL, *Multimodal Virtual Point 3D Detection*; in proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), Electr Network, F Dec 06-14, 2021 .2021.

T.Liang, H.Xie, K.Yu, Xia, et al, ArXiv (2022).

X. Y.Bai, Z. Y.Hu, X. G.Zhu, et al, *TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers*; in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, F Jun 18-24,(2022).

Y. W., A W., T J.Meng, et al, *DeepFusion: Lidar-Camera Deep Fusion for Multi-Modal 3D Object Detection*; in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, F Jun 18-24,(2022).

Y.Li, Y.Chen, et al, ArXiv (2022).

Y.Ma, T.Wang, X.Bai, H.Yang, ArXiv.2208.02797(2022).

Y.Zhou, O.Tuzel, *VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection*; in proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, F Jun 18-23, (2018).

Z. B. Huang, S L.Sun, J.Zhao, et al, Inf Fusion, 98: 11(2023).

Z. Q.Li, et al, *BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers*; in proceedings of 2022 European Conference on Computer Vision(ECCV).

Z. Y.Chen, et al, ArXiv.2207.10316. Accessed 3 Feb. (2024).

Z. Y.Chen, et al, *AutoAlign: Pixel-Instance Feature Aggregation for Multi-Modal 3D Object Detection*; in Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence(IJCAI), 1 July 2022, ijcai.2022/116. Accessed 3 Oct. (2023).

Z.Chen, Z.Li, S.Zhang, et al, ArXiv,(2022).