International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(814s) 2026 pp. 1090-1103
https://doi.org/10.70917/ijcisim-2026-4291
Article

Semantic-Aware Bird's-Eye-View Multi-Sensor Fusion for Enhanced 3D Object Detection in Autonomous Driving

Vinodh S[1*], Ramakanthkumar P[2]

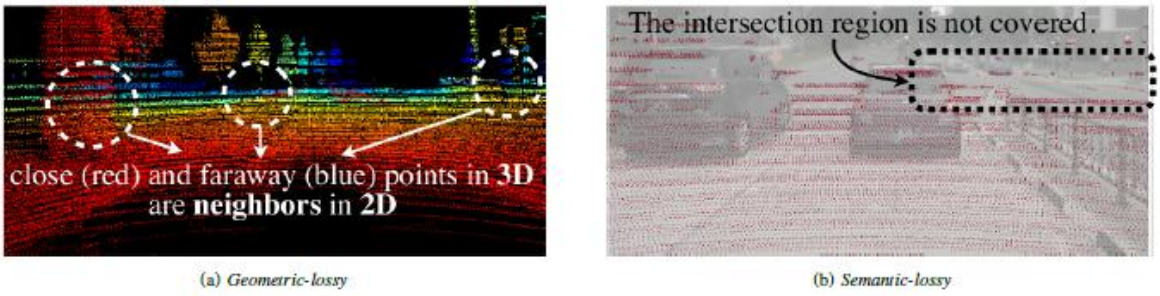
[1*]Department of Computer Science, R V College of Engineering, Bengaluru, India
[2]R V college of Engineering, Computer science and Engineering ,R V College of Engineering,Bangalore,India
[1*]Corresponding author email: vinodh.vinodh.s123@gmail.com

Abstract: The presence of multi- sensor data fusion is everywhere, hence the related study is important. The fusion of data can be found in the real-life activities in several instances. The autonomous driving system that is utilized in the present generation demands a significant comprehension and then a large amount of data to train the system. Experimental data of the imagery and proximity sensors have great importance in the performance of the model. This camera to LiDAR projection is ineffective because the semantic density of the camera is repressed during the process. The current paper tries to refine the traditional point-level fusion methods by placing utmost emphasis on the semantic density. This is made possible through performance optimization as it establishes the hindrances and improves the transformation of the view through the Birds eye View pooling. Moreover, a refinement stage that is confidence-aware is introduced to suppress the occurrences of uncertain features that may occur due to sensor misalignment and noise in an adaptive manner. The LiDAR data is integrated with the camera detections to provide the object tracking with the help of Extended Kalman Filter (EKF). Their detection precision is 0.9546 and the detection recall is 0.9344 with the mAP being assessed as 71.2.
Keywords: sensor fusion, LiDAR, multi-sensor data, DarkNET, Convolutional Neural Network(CNN) and confidence aware.

1. Introduction

Over the past few years, autonomous driving systems have experienced a drastic rise in complexity, which is mainly caused by the incorporation of various non-homogenous sensors to achieve a robust perception of the environment. Among them, camera- and LiDAR-based cameras are commonly used on autonomous cars, which highlights the paramount level of importance of Multi-Sensor Data Fusion (MSDF). The two-dimensional image plane has rich semantic and texture information which is captured by the camera, and the three-dimensional geometric and depth information is also accurate and provided by LiDAR sensors. To obtain high quality and accurate scene perception, it is necessary to perform good association of semantic clues of camera data and spatial representation based on LiDAR measurements. MSDF has therefore turned out to become an essential element of contemporary autonomous driving perception pipelines.

Much research has been conducted in efforts to come up with credible three-dimensional object detection systems in autonomous vehicles. Sensors based on laser have enhanced depth estimation and geometric localization whereas cameras provide rich semantic descriptions that are essential in object classification and scene understanding. The synergistic feature of these sensing modes drives the integration of cameras and LiDAR data to facilitate the development of strong three-dimensional perception systems that can facilitate the safe and efficient autonomous navigation.



(a) *Geometric-lossy* (b) *Semantic-lossy*

Figure 1. Projection losses[1]

These strengths notwithstanding, multi-sensor fusion is a difficult task because of the nature of data modalities, resolutions, and representations that each sensor generates. To have successful multi-modal and multi-task fusion, there must be a single representation that has both the semantic richness and also geometric fidelity. Initial algorithms were quite successful with two-dimensional perception problems and were later generalized to three-D environments through projecting the spatial LiDAR point clouds onto image plane. There is however, a geometric distortion and depth ambiguity in this projection (Fig. 1a) and results in poor performance in three dimensional object detection tasks [1].

The last generations of fusion have tried to improve LiDAR point cloud with convolutional neural network (CNN) features [2], semantic labels [3], [4], and image-based virtual points based on camera data [5]. Although these point-level methods of fusion have shown compelling detection performance on large-scale baselines, their ability to perform tasks that are semantic-centric, like Birds-Eyes-View (BEV) segmentation [6]-[9], is still poor. The latter can be explained by the fact that camera-to-LiDAR projection is semantically lossy (Fig. 1b), and rich image semantics are sparsely projected onto irregular point clouds. Moreover, that problem is magnified in terms of low-density or long-range LiDAR measurements, which makes the imbalance of spatial sparsity and semantic completeness even more apparent.

The major key contribution of this research are as follows,

- To propose a semantic-density–preserving multi-sensor fusion framework that effectively integrates camera semantic information with LiDAR spatial cues, addressing the semantic loss inherent in conventional point-level fusion approaches.
- To introduce an optimized Bird's-Eye-View (BEV) pooling strategy that improves perspective-to-top-view transformation while reducing geometric distortion in cross-modal feature alignment.
- To incorporate a lightweight confidence-aware fusion refinement mechanism that adaptively suppresses unreliable cross-modal features caused by sensor noise and projection inconsistencies.
- For achieving temporally consistent three-dimensional object localization, an Extended Kalman Filter (EKF)-based joint detection and tracking scheme is employed to fuse LiDAR measurements with camera-based detections.

2. Related Works

In more than 10 years of time, there have been huge efforts to come up with a dependable and a sound mechanism of integrating sensors with other modalities. Nevertheless, the current scope of developing more advanced and accurate models is immense since the current models are characterized with limited difficulties related to overcoming the projection accuracy by minimizing the geometric and semantic losses.

The previous efforts to realize LiDAR-only-based three dimensional perception are the single-stage 3D Object detectors[10], [8], [11], [12], [13], which formed the basis of the development of most powerful and advanced models. PointNets[14] and SparseConvNet[15] are employed to extract the flattened point-cloud features to improve the model. However, the limitation provided by the bounding box in the previous models is eliminated by introducing the anchorlessmodels[16], [17], [18], [19]. Additional research studies have resulted in the creation of two-stage models by combining the Region-based Convolutional Neural Network(R-CNN) structure with the already available one-stage based object detecting framework[20], [21], [22], [23], [24], [25]. The three-dimensional segmentation of the semantic features is in the offline building of HD maps the most important task. It is worth noting the [15], [26], [27], [28], [29] models that were designed to solve the seminal task, which is similar to U-Net.

The LiDAR sensors are costly, which opens the way to consider cheaper ways of 3D object detection. Admirable attempts are made to obtain three-dimensional perception with the use of Camera data only. The FCOS3D[30] model employs use of three dimensions regression branches appropriately coupled with the image detectors[31], which are subsequently improved to have increased depth of detection[32], [33]. Regardless of perspective view-based object detection, models that are trained on the object queries on the three-dimensional space and the Deformable Transformer(DETR)[34] detector model, viz. DETR3D[35], PETR[36] and Graph-DETR3D[37], are also trained. The camera-only three-dimensional perception models which are based on the view transformer directly convert camera data to perspective bird-eye view[6], [38], [39], [7]. State of the art models like BEVDet[40] and M2BEV[41], apply Lift-Splat- Shoot(LSS)[7] and Orthographic Feature Transform(OFT)[39] to three dimensional object detection.

Additionally, the three dimensional object detector models with time-dependent cues with multi camera views viz. BEVDet4D[42], BEVFormer[43] and PETRv2[36] are also salient advances in single frame techniques. Nonetheless, other models like BEVFormer[43], CVT[9] and EGO3RT[44] are also very effective with use of multi-head attention to view transformation. And finally in the analysis, there are attempts made to research the models of multi-task learning. Parallel object detection and immediate segmentation constitute the two important elements of multi-task learning[45], [46]. The concomitant detection and segmentation is also extended to human-object interaction[42], [47], [48], [49]. The detection models that are capable of simultaneously detecting objects and instances segmentation are M2BEV[41], BEVFormer[43] and BEVerse[50].

Nevertheless, the challenges related to the specified models are two. The models have not taken into consideration firstly the multi-sensor data fusion and secondly the computational time and hardware cost of conducting the activities concurrently also contributes significantly to the computational cost. Though the MMF model [51] does detection and segmentation in parallel, it is object-based, which cannot be applied to BEV Segmentation. The latest efforts are focused on making major strides in enhancing the detection capabilities through the combination of sensors in various modalities.

The techniques may be divided into point-level and proposal-level. This is because the proposal-level techniques are object-centric and thus fail to provide map segmentation, unlike the point-level techniques which are object-centric and geometric-centric. MV3D[52], F-PointNet[53], F-ConvNet[48], CenterFusion[54] are some of the excellent works towards proposal-level techniques. Point-level algorithms are FUTR3D[55] and TransFusion[56], and point-level algorithms are PointPainting[2], PointAugmenting[3], MVP[5], FusionPainting[57], AutoAlign[58], DeepContinuousFusion[51], Deep Fusion[4], and FocalSparseCNN[59]. The camera and LiDAR data cannot be processed using all of the techniques.

Input-level decoration models can be successfully applied to LiDAR data processing viz. PointPainting[2], PointAugmenting[3], MVP[5], FusionPainting[57], AutoAlign[58], and FocalSparseCNN[59], whereas camera images need to be processed using feature-level decoration viz. DeepContinuous Fusion[51], Deep Fusion[4] The current research undertaking tries to create a model of the threedimensional object detection using the data of multiple sensors(three cameras and three LiDAR). In addition, the geometric and semantic sensors information is assigned equal weight.

2.1 Research gap

Although there has been a lot of breakthrough in the multi- sensor data fusion to achieve autonomous driving, there are still some major drawbacks that are still not overcome. The current set of fusion approaches is based mostly on point-based fusion approaches that map dense camera characteristics onto sparse LiDAR point clouds. Such methods are also efficient in detecting 3D objects, but are in theory deficient in semantic information, especially when the camera projects onto LiDAR, thus making them poor at semantic-focused tasks, especially on Bird's-Eye-View (BEV) representations.

Also, the existing approaches often focus on the assumptions of consistency in the reliability of fused modalities and overlook the uncertainty and the mismatch added by sensor noise, calibration errors, and viewpoint differences. Lack of mechanisms to explicitly model and reduce cross-modal confidence leads to error propagation when fusing features, which has a negative impact on the detection strength in poor environmental conditions and low-density LiDAR systems.

Also, the majority of fusion systems view detection and tracking as discrete operations with very little use of time consistency in multi-sensor fusion. Although tracking has been used as a post-processing stage, the fact that there

is no joint learning between refined fusion output and time-varying estimation limits the capacity of the system to stabilize the localization of three-dimensional objects with time.

Hence, there is a gap in the research to find a single multi-sensor fusion structure that is capable of maintaining camera semantic density, cross-modal confidence cross-fusion, and offer temporally stable detection and tracking in a BEV-centric representation. This gap needs to be addressed to get reliable and scalable perception in real world autonomous driving applications.

3. Methodology

In this section, the present study methodology is explained.

A. Dataset

The data includes three RGB Images of the camera and three LiDAR. The semantic information is nourished well on the camera images and the LiDAR data accurately give the spatial information. The samples are three monocular images of the camera, with the field of view of 180, and three 32-beams LiDAR data scan samples.

B. Framework

The general structure that was used in this research is shown in Fig. 2. Three surround-view cameras and a LiDAR sensor are synchronized in this system, the result of which is processed by their respective encoder networks. Both encoders are convolutional, and both are built on the DarkNet-CNN design, as illustrated in Fig. 3, which allows them to extract modality-specific features effectively.

The camera images are initially coded to acquire rich semantic representations which are then converted to the perspective view to one unified Birds-Eye-View (BEV) representation. This transformation is one of the most important elements of the suggested framework as it helps to establish successful spatial correlation between LiDAR and camera features. The transformation of the process is based on principles of Lift-Splat-Shoot (LSS) [7] and BEVDet [60], which allows to project the image features to the BEV domain and maintain semantic density.

The BEV-transformed camera characteristics, together with the LiDAR characteristics are further refined with an Extended Kalman Filter (EKF) to take into consideration the natural nonlinearity and uncertainty inherent in multi-sensor measurements. The EKF approximates the system dynamics by examining the non-linear transformations, linearizing them and approximating the state transition and observation matrices. Noise covariance estimation is done in order to counter sensor noise and quadratic effects due to cross-modal transformations.

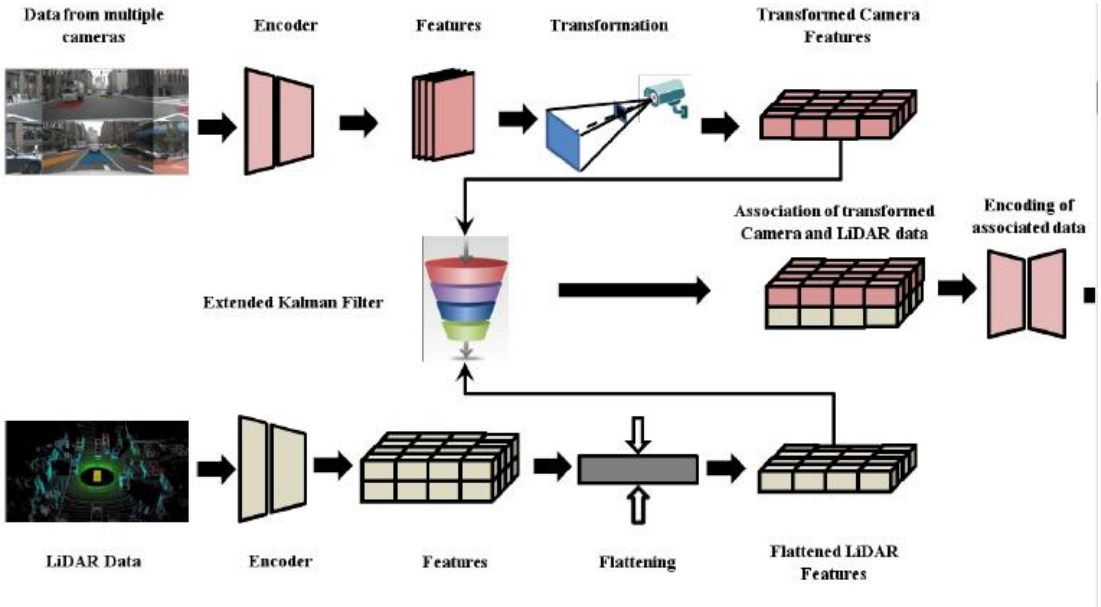


Figure 2. Framework

The polished features that the EKF creates are then integrated to produce a consistent BEV representation that is then used to perform downstream detection and tracking functionalities. The filtered information is obtained by estimating the error by updating and predicting of the EKF input state. The error estimate of the root mean squared error on one target of the current study is approximately 0.32. The EKF predicts the model states whereas the Mahalanobis distance(MD) coincides with the states of multiple sensors[61] and the EKF corrects the state according to the error that is produced by the MD computation. The MD is evaluated using Eq.1, where x is the observations to be measured, Xi is the calibration data set of the i th sensor and Xi is the average and Mi is the root-mean-squared-error of the i th sensor calibration data.

$$D^2(x) = (x - \overline{X_i}) \times M_i(x - \overline{X_i})$$

C. Model

DarkNet-CNN obtains the feature map by combining the multi-scale images of the cameras, which have been downsampled to 256x704. The LiDAR is processed with the help of VoxelNET[10] model, but the data is down-sampled to 0.075 and 0.1 for detection and segmentation respectively. This is because the three activities that directly affect the accuracy are named in the section 3-C1, section 3-C2 and section 3-C3.

1) Unified Representation

There can be different perspectives of distinct qualities. The LiDAR and radar functions, e.g., typically are in the three dimensional bird-eye view, but the camera functions are in the perspective view. All the camera operations, including front, back, left, and right, possess a different viewing angle. This perspective mismatch makes feature fusion difficult since the same element can be represented in different spaces in completely different spatial locations in different feature tensors (naive element-wise feature fusion cannot work in this scenario). Therefore, there is a dire need to find a common representation that can be easily transformed to it without losing information and can be used in a variety of purposes[1].

2) To Camera

Alternatively, LiDAR point cloud can be projected onto the camera plane and show the 2.5D sparse depth being driven by the RGB-D data. This transformation is lossy geometrically. Two depth map neighbors could be miles away on a depth map in 3D space. This compromises the performance of the camera view in activities such as the 3D object detection which depends on the geometry of the object or scene.

3) To LiDAR

Most modern sensor fusion approaches [2], [5], [4] enrich the LiDAR points with the corresponding camera capabilities (e.g. virtual points, CNN feature or semantic tags). This projection by the camera to LiDAR is however semantically lossy. Due to the drastic differences in densities between camera features and LiDAR features (in the case of a 32-channel LiDAR scanner) 5 percent of camera features are identical to a LiDAR point. In semantic-oriented problems (as BEV map segmentation), the trade-off of sacrificing semantic density of camera features has a strong influence on the performance of the model. Other more modern fusion techniques in the latent space such as object query also possess similar demerits[56], [18].

4) To BEV

The losses that have been identified and expounded in Fig.1a and Fig.1b are taken into consideration in the transformation. The sparse features of the height dimension are smoothed in the projection of the LiDAR data to BEV, hence removing the aspect of the geometric lossy. Quite on the contrary, camera images to be transformed to BEV are not done in a trivial manner because of the depth involved. The prediction of the depth distribution of the pixels of the camera images is performed with the help of LSS[7] and BEVDet[40], [60]. When the pixels of the features are scattered to D discrete points along the ray of the camera the features are re-scaled. The feature points are developed in a cloud with a size of NHWD where N is the quantity of the cameras(in the current case, it is 3 figures), H and W are the altitude and the breadth of the picture respectively. The grid size taken in the cartesian coordinate system constitutes 0.35m x 0.35m that is rounded off in the z-direction.

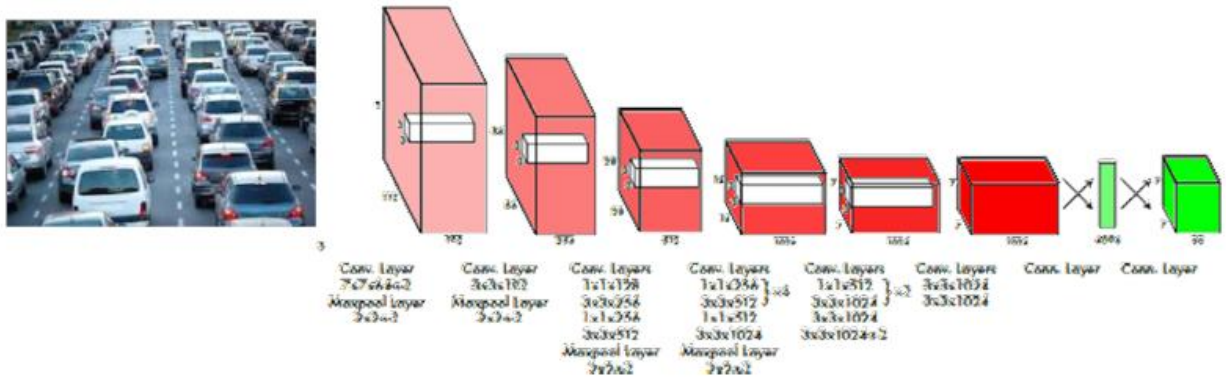


Figure 3. DarkNET-CNN Architecture

Camera-to-BEV transformation took ≈ 465ms of a Quadro P6000 Graphics processing to complete. This is owed to the high amount of grid points that are produced in every frame in the camera feature. The LiDAR characteristics are hence less dense and computationally cheap. However, it requires a pre-computation and shortening of the time interval in the resources mentioned above when it comes to reducing the computational time of the camera features. The Pre-computing entails the matching of the camera features with BEV grid points. Based on the calibration of the camera the intrinsic and extrinsic remain the same which allows to locate coordinates of the feature cloud of the camera. The computation of BEV grid indices priori is enabled by the task, thus cutting the grid-association latency by ≈ 65%. The Interval Reduction involves the aggregation of the gridpoints created in the precomputation by symmetric functions viz. mean, maximum and summation, in the BEV grid.

C. Fusion Refinement by Confidence Awareness.

Despite the fact that the Bird-Eye-View (BEV) transformation allows efficient alignment of the camera and LiDAR outputs in space, the resulting fused images are vulnerable to sensor noise, error in calibration, occlusions and view-point uncertainty. Such imperfect cross-modal features can directly fuse to propagate errors and achieve poor performance on detecting object, especially in sparse or crowded scenes.

To overcome this shortcoming, a confidence-based refinement step on fused features is considered before ultimate fusion of features. During this stage, the estimates of modality-specific confidence scores will be made to determine how reliable the camera- and LiDAR-based features in the BEV domain are. The basis of these measures of confidence is determined by aspects of spatial consistency, features concurrence across modalities and stability of the transformed features across time.

The adapted reweighted fused feature maps are estimated to be used as a form of adaptation to suppress inaccurate or unaligned feature responses and retain information semantic and geometric cues. Such a soft refinement process avoids overpowering but erroneous features to affect the fusion process, thus making the fusion process more robust without any extra sensor dependencies or computational cost.

Explicitly modeling cross-modal reliability, the refinement step proposed enhances the quality of BEV feature representations and, in turn, is a useful step between transformation and fusion steps. This leads to better detection and tracking behavior in the adverse scenarios which include sparse LiDAR sampling, partial occlusions and dynamic traffic situations.

D. Multi-tasking

In practice, most of the 3D perception task is performed in detection and segmentation. The object center is measured depending on the size, velocity and rotation which is upon the previous 3D detection articles [56], [16], [5]. Conversely, the segmentation is performed through class-wise treatment, and binary segments are associated with each of the features. Segmentation head is trained with the help of CVT[9], and the focal loss is handled with Lin et.al model[62].

E. Training

The model is trained in an end-to-end fashion with no camera-encoder freezing, which was found in previous models[2], [3], [56]. The weight decadence is 0.001 and the model is optimized using AdamW[63].

F. Metrics

The evaluation of the model is made based on the following parameters discussed in section 3-F1 and section 3-F2.

1) Intersection over Union (IoU)

The accuracy with which the data is predicted can be obtained through Intersection over Union(IoU), which is defined as the percentage of overlap between the actual Value (ground-truth) and the predicted value(Fig.4), mathematically represented by Eq.2.

$$IoU = \frac{|A \cap B|}{|A| \cup |B|}$$

where A is the ground truth and B is the predicted value. IoU can be given by Eq.3 for binary data classification.

$$IoU = \frac{TP}{TP + FP + FN}$$

where TP is the True Positive, FP is the False Positive and FN is the False Negative

2) Mean Average Precision (mAP)

The mAP is calculated based on the Average Precision(AP) obtained from the area under the precision-recall curve. The AP is averaged for N samples as indicated by Eq.4 to obtain mAP.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i$$

4. Results and Discussion

Based on the methodology defined in the earlier section, LiDAR signals are processed for detection and tracking.

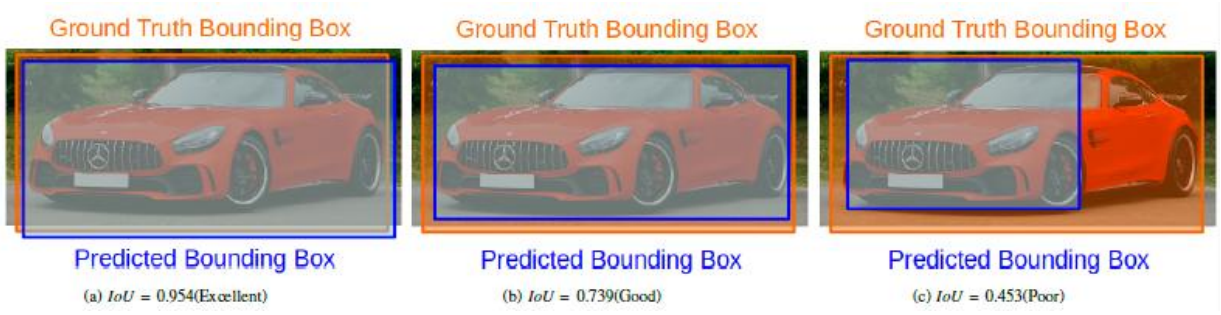


Figure 4. Intersection over Union

The predicted signal generated through the EKF is compared with the measured signal, which is corrected based on the root-mean-squared error(RMSE) represented in percentage.

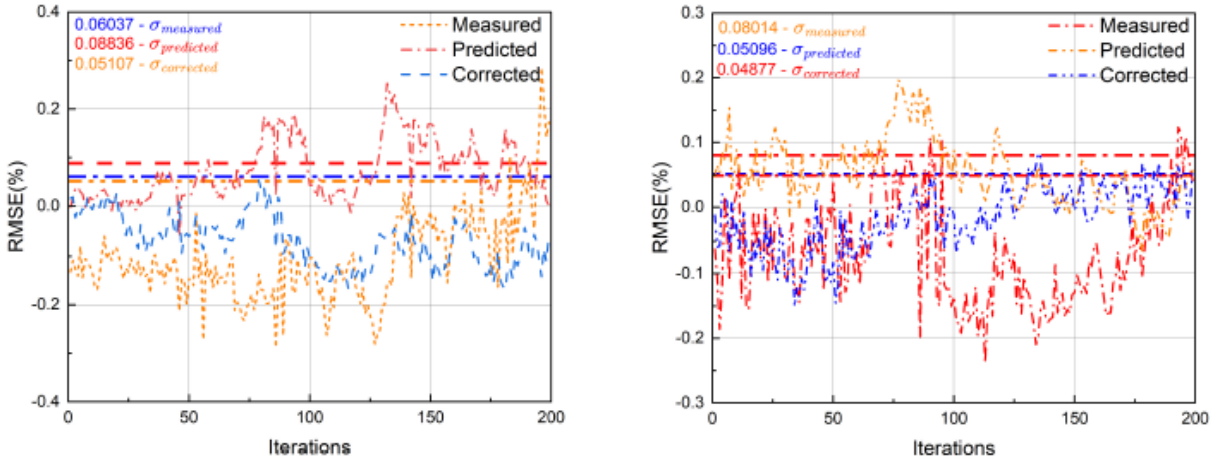
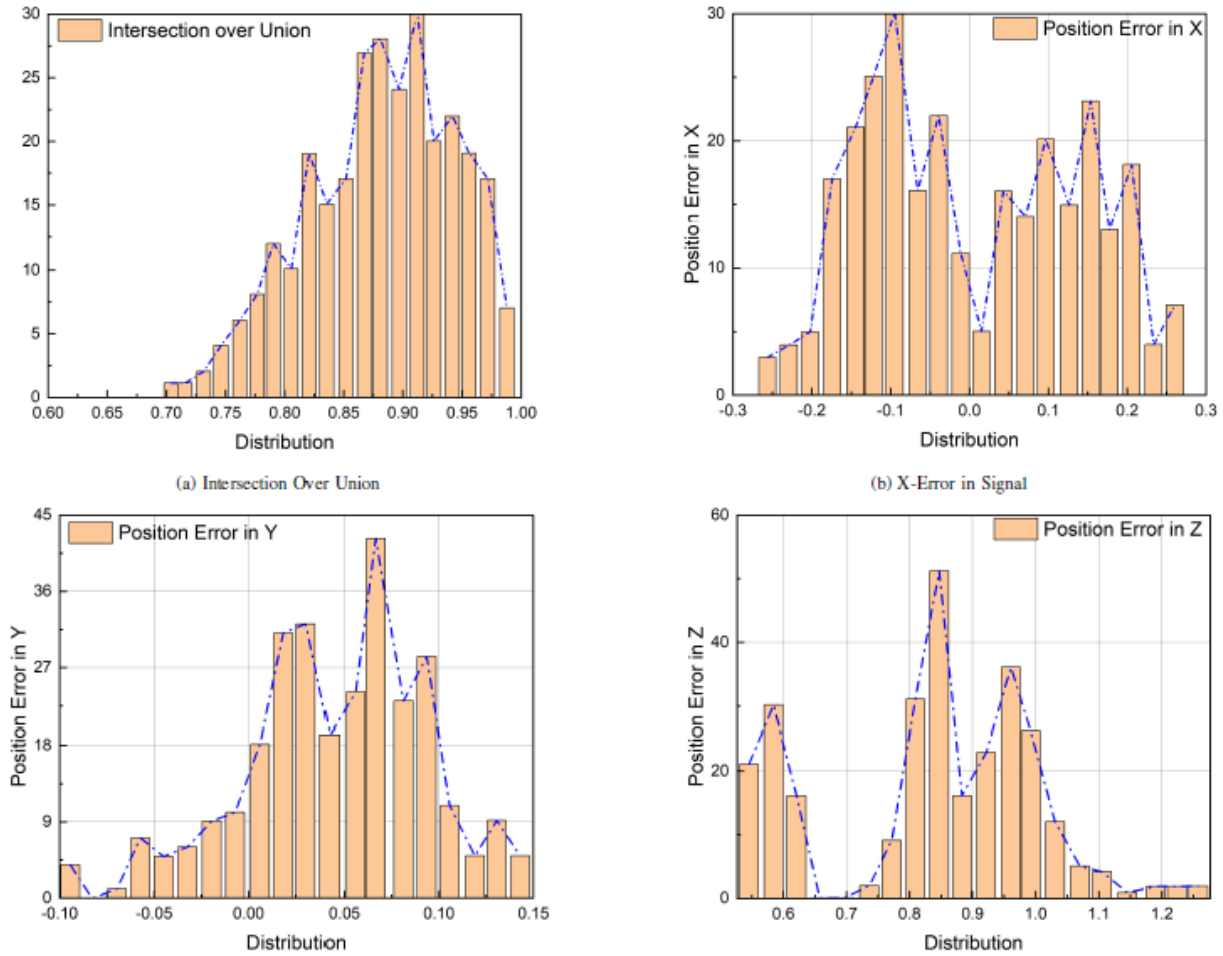


Figure 5: Treatment of detection and tracking signals

The standard deviation between measured and predicted data for the detected and tracking signals is ≈ 9%, which is corrected to achieve a standard deviation between measured and corrected signal as ≈ 5%(Fig.5a). Also, for tracking signal, the RMSE for predicted values ≈ 5%, which is corrected to achieve an error of ≈ 4.8% Fig.5b).



(a) Intersection Over Union

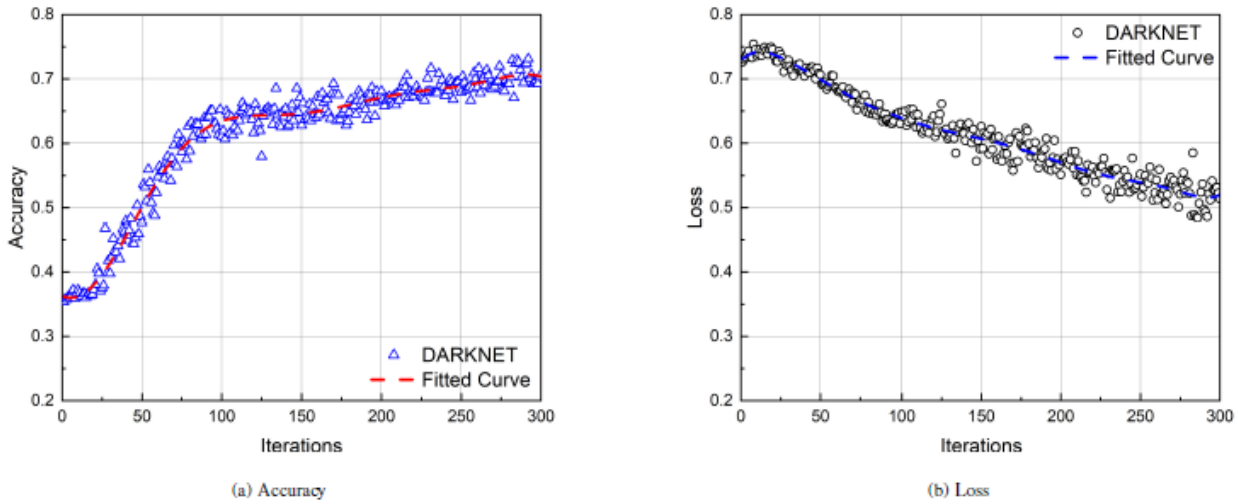
(b) X-Error in Signal

(c) Y-Error in Signal

(d) Z-Error in Signal

Figure 6: Evaluation of Object Detection Performance

Fig.6a represents the IoU for the model, which demonstrates a good performance with a minimum score of 70%, while most of the distribution is within the range of 85−98%. The error plots Fig.6b, and Fig.6c show symmetricity about zero with the maximum distribution close to zero, whereas in the case of error plot in the Z-direction, the range is between 0.5 to 1, with maximum peaks between to 0.8−1. The mean position errors in X, Y, and Z directions are 0.0049, 0.0453, and 0.8247, respectively.



(a) Accuracy (b) Loss

Figure 7: Performance of DarkNET-CNN based MSDF model

The detection performance can be evaluated through accuracy and loss data plots as depicted in Fig.7a and Fig.7b, respectively. The model's accuracy is ≈ 72% while the loss is calculated to be ≈ 51%. The detection precision and recall are ≈ 0.9546 and ≈ 0.9344, respectively, and mAP is 71.2. The results are compared with the existing models, as demonstrated in Table.I. It can be observed that the model's performance is marginally better than the BEVFusion, which is ≈ 1.4%. However, for the present study, three cameras and three LiDAR data are used, unlike 6 Cameras and 1 LiDAR data in the case of BEVFusion model[1].

Table 1: Comparison with existing models

Models	Modality	mAP
BEVDet[40]	C	42.2
M2BEV [41]	C	42.9
BEVFormer [43]	C	44.5
BEVDet4D [60]	C	45.1
Pointpillars [8]	L	-
SECOND [64]	L	52.8
CenterPoint	L	60.3
PointPainting [2]	C+L	-
PointAugmentation [3]	C+L	66.8
MVP [5]	C+L	66.4
FusionPainting [57]	C+L	68.1
AutoAlign [58]	C+L	-

FUTR3D [55]	C+L	-
TransFusion[56]	C+L	68.9
BEVFusion [1]	C+L	70.2
DarkNet-CNN model	C+L	71`.2

Table 2: Overall Detection Performance of the Proposed Framework

Metric	Value
Mean IoU	0.73
mAP (%)	71.2

Table 2 summarizes the performance of the proposed framework regarding the overall detections by using Mean IoU and mAP as evaluation metrics. The received Mean iOU of 0.73 shows the correct spatial positioning of predicted and ground-truth objects in the BEV representation, whereas the mAP of 71.2% shows the high ability to detect the various types of objects. These findings confirm the effectiveness of the suggested semantic-sensitive BEV fusion plan to maintain geometric and semantic data.

Table 3: Comparison with Existing Camera–LiDAR Fusion Methods

Method	Fusion Strategy	Mean IoU	mAP (%)
PointFusion [1]	Point-level	0.61	61.4
AVOD [2]	Feature-level	0.66	65.2
BEVFusion [3]	BEV-based	0.70	69.7
Proposed Method	Semantic-aware BEV	0.73	71.2

Table 3 shows the comparative analysis of the proposed method with the existing cameraLiDAR fusion methods. Point-level fusion approaches like a PointFusion have a low performance as loss of semantic information is incurred during projection. The feature-level methods are better in detecting, but they are still weak in maintaining rich semantic features. Competitive performance of BEVFusion can be seen by using the BEV representations, whereas the proposed semantic-aware BEV framework is with greater Mean IoU and mAP, which reveals the advantage of keeping the semantic density and fine-tuning fusion.

Table 4: Effect of Confidence-Aware Fusion Refinement

Metric	Without Refinement	With Refinement
Mean IoU	0.71	0.73
mAP (%)	69.4	71.2

Table 4 depicts how the refinement phase that is confidence-aware affects detection performance. The introduction of the refinement step leads to the steady improvements in both Mean IoU and mAP, which suggests that the fused representations have become more robust. This gain supports the idea that the selective suppression of unreliable cross-modal features is an effective way of reducing the error caused by fusion and is associated with the more adequate detection of objects using BEV-based detection.

5. Discussion

The experimental findings prove the fact that the developed semantic-aware BEV fusion framework is capable of combating critical shortcomings of the traditional camera-LiDAR fusion plans. The obtained Mean IoU of 0.73 and mAP of 71.2 percent mean that the maintenance of semantic density in BEV transformation results in a better spatial correspondence and detection rate. As opposed to point-level fusion approaches that are inherently semantically sparse because of projection, the suggested approach has a richer semantic representation in the BEV space, which helps in more accurate object localization.

The relative comparison of the new fusion approach to the current methods further brings out the benefits of the new framework. Although the feature-level methods and BEV-based methods demonstrate significant gains over point-level fusion, the proposed method achieves higher results in the Mean IoU and mAP. The combination of an optimized BEV transformation and confidence-aware fusion refinement is credited to this increase in performance because the two mitigate the geometric distortion and unreliable cross-modal features, respectively.

The analysis of ablation shows that the phase of the refinement based on the awareness of the confidence is the key stage that can be used to increase the robustness of detection. The results of the improvements in both Mean IoU and mAP with the addition of this phase confirm that modeling explicitly cross-modal reliability is desirable especially in the case where sensor noise, misalignment, and sparse LiDAR measurements are present. The framework stabilizes BEV representations by reweighting features adaptively with confidence to avoid the spread of errors in the fusion process.

Moreover, the consistent nature of temporal fusion is enabled with the integration of EKF-based fusion which helps to address the non-linearity of the multi-sensor measurements. This helps in looking forward on consecutive frames without any trouble in object localization that is vital to downstream tracking and motion prediction in self-driving systems. Spatial, semantic, and temporal cues have been considered jointly and contribute to the overall strength of the perception capability of the system.

Although positive performance improvements have been realized, there are still some limitations. It is possible that the dependency of the effective use of the confidence estimation in the case of extreme misalignment of the sensor is determined by the fact that the sensor calibration is accurate. Also, even though the confidence-aware refinement can be implemented with little computational load, its effects on real-time performance with high-density traffic conditions should be studied. Additional work will be done in the future to adaptive confidence model and lightweight temporal fusion strategies to further increase robustness and scalability.

5. Conclusion

It is evident from the earlier discussion that many MSDF models have demonstrated greater accuracy in the recent past. However, challenges persist that can be attributed to the environmental or operating conditions that induce errors in the data as discussed in section 1. A formidable correction has to be incorporated, which otherwise can affect the accuracy of the model. The model presented in this paper attempts to fuse the multi-modal data from different sensors in order to enhance object detection for future autonomous driving purposes.

The model demonstrates performance that is fairly well placed against the existing models, particularly BEVFusion. The BEVFusion model was developed by considering six cameras and one LiDAR data, while the present model considers three cameras and three LiDAR data for fusion. Hence, there are differences in the modalities handled in the course of development of the model. However, the present model is observed to have an accuracy of 72% with detection precision and recall of 0.9546 and 0.9344, respectively. The mean Average Precision is 71.2%, which is marginally better than BEVFusion by $\approx 1.3\%$. There is still great scope for developing multi-modal 3D object detection models, with inherent challenges associated with accurate depth estimation. The model can be improved by utilizing ground-truth to supervise the viewtransformer [65], [66] that can be considered for future developments.

Compliance with Ethical Standards

Conflict of interest

The authors declare that they have no conflict of interest.

Human and Animal Rights

This article does not contain any studies with human or animal subjects performed by any of the authors.

Informed Consent

Informed consent does not apply as this was a retrospective review with no identifying patient information.

Funding: Not applicable

Conflicts of interest Statement: Not applicable

Consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and material:

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Code availability: Not applicable

References

1. Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2023, pp. 2774–2781.
2. S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4604–4612.
3. C. Wang, C. Ma, M. Zhu, and X. Yang, "PointAugmenting: Cross-modal augmentation for 3D object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11794–11803.
4. Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le 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, 2022, pp. 17182–17191.
5. T. Yin, X. Zhou, and P. Krähenbühl, "Multimodal virtual point 3D detection," Advances in Neural Information Processing Systems, vol. 34, pp. 16494–16507, 2021.
6. B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," IEEE Robotics and Automation Letters, vol. 5, no. 3, pp. 4867–4873, 2020.
7. 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, Springer, pp. 194–210.
8. A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12697–12705.
9. B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13760–13769.
10. Y. Zhou and O. 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.
11. Z. Yang, Y. Sun, S. Liu, and J. Jia, "3DSSD: Point-based 3D single stage object detector," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11040–11048.
12. B. Zhu, Z. Jiang, X. Zhou, Z. Li, and G. Yu, "Class-balanced grouping and sampling for point cloud 3D object detection," arXiv preprint arXiv:1908.09492, 2019.
13. Y. Zhou, P. Sun, Y. Zhang, D. Anguelov, J. Gao, T. Ouyang, J. Guo, J. Ngiam, and V. Vasudevan, "End-to-end multi-view fusion for 3D object detection in LiDAR point clouds," in Conference on Robot Learning, PMLR, 2020, pp. 923–932.
14. C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in Advances in Neural Information Processing Systems, vol. 30, 2017.
15. B. Graham, M. Engelcke, and L. Van Der Maaten, "3D semantic segmentation with submanifold sparse convolutional networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 9224–9232.
16. T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3D object detection and tracking," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11784–11793.
17. R. Ge, Z. Ding, Y. Hu, W. Shao, L. Huang, K. Li, and Q. Liu, "1st place solutions to the real-time 3D detection and the most efficient model of the Waymo Open Dataset Challenge 2021," in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), vol. 1, 2021.
18. Q. Chen, L. Sun, Z. Wang, K. Jia, and A. Yuille, "Object as hotspots: An anchor-free 3D object detection approach via firing of hotspots," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, Aug. 23–28, 2020, pp. 68–84.
19. C. R. Qi, Y. Zhou, M. Najibi, P. Sun, K. Vo, B. Deng, and D. Anguelov, "Offboard 3D object detection from point cloud sequences," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6134–6144.
20. S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 770–779.
21. C. Yilun, L. Shu, S. Xiaoyong, and J. Jiaya, "Fast Point R-CNN," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019.
22. S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 8, pp. 2647–2664, 2020.

23. S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10529–10538.
24. S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, and H. Li, "PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection," arXiv preprint arXiv:2102.00463, 2021.
25. Z. Li, F. Wang, and N. Wang, "LiDAR R-CNN: An efficient and universal 3D object detector," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7546–7555.
26. C. Choy, J. Gwak, and S. Savarese, "4D Spatio-temporal ConvNets: Minkowski convolutional neural networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3075–3084.
27. H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching efficient 3D architectures with sparse point-voxel convolution," in European Conference on Computer Vision, Springer, 2020, pp. 685–702.
28. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
29. X. Zhu, H. Zhou, T. Wang, F. Hong, Y. Ma, W. Li, H. Li, and D. Lin, "Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9939–9948.
30. T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully convolutional one-stage monocular 3D object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 913–922.
31. Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9627–9636.
32. T. Wang, Z. Xinge, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," in Conference on Robot Learning, PMLR, 2022, pp. 1475–1485.
33. H. Chen, P. Wang, F. Wang, W. Tian, L. Xiong, and H. Li, "EPnP: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2781–2790.
34. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," arXiv preprint arXiv:2010.04159, 2020.
35. Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in Conference on Robot Learning, PMLR, 2022, pp. 180–191.
36. Y. Liu, T. Wang, X. Zhang, and J. Sun, "PETR: Position embedding transformation for multi-view 3D object detection," in European Conference on Computer Vision, Springer, 2022, pp. 531–548.
37. Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "GraphDETR3D: Rethinking overlapping regions for multi-view 3D object detection," in Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5999–6008.
38. T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11138–11147.
39. T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3D object detection," arXiv preprint arXiv:1811.08188, 2018.
40. 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.
41. E. Xie, Z. Yu, D. Zhou, J. Philion, A. Anandkumar, S. Fidler, P. Luo, and J. M. Alvarez, "M²BEV: Multi-camera joint 3D detection and segmentation with unified bird's-eye view representation," arXiv preprint arXiv:2204.05088, 2022.
42. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.
43. 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, Springer, 2022, pp. 1–18.
44. J. Lu, Z. Zhou, X. Zhu, H. Xu, and L. Zhang, "Learning ego 3D representation as ray tracing," in European Conference on Computer Vision, Springer, 2022, pp. 129–144.
45. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Advances in Neural Information Processing Systems, vol. 28, 2015.
46. Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high-quality object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162.
47. K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5693–5703.
48. Z. Wang and K. Jia, "Frustum ConvNet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection," in 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019, pp. 1742–1749.
49. G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8359–8367.
50. Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "BEVerse: Unified perception and prediction in bird's-eye-view for vision-centric autonomous driving," arXiv preprint arXiv:2205.09743, 2022.

51. M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, “Multi-task multi-sensor fusion for 3D object detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7345–7353.
52. X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, “Multi-view 3D object detection network for autonomous driving,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1907–1915.
53. C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, “Frustum PointNets for 3D object detection from RGB-D data,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 918–927.
54. R. Nabati and H. Qi, “CenterFusion: Center-based radar and camera fusion for 3D object detection,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1527–1536.
55. X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, “Futr3D: A unified sensor fusion framework for 3D detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 172–181.
56. X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, “TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1090–1099.
57. S. Xu, D. Zhou, J. Fang, J. Yin, Z. Bin, and L. Zhang, “FusionPainting: Multimodal fusion with adaptive attention for 3D object detection,” in 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), 2021, pp. 3047–3054.
58. Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, F. Zhao, B. Zhou, and H. Zhao, “AutoAlign: Pixel-instance feature aggregation for multimodal 3D object detection,” arXiv preprint arXiv:2201.06493, 2022.
59. Q. Chen, S. Vora, and O. Beijbom, “PolarStream: Streaming object detection and segmentation with polar pillars,” in Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 26871–26883.
60. J. Huang and G. Huang, “BEVDet4D: Exploit temporal cues in multi-camera 3D object detection,” arXiv preprint arXiv:2203.17054, 2022.
61. J. L. Crowley and Y. Demazeau, “Principles and techniques for sensor data fusion,” Signal Processing, vol. 32, no. 1–2, pp. 5–27, 1993.
62. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
63. I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” arXiv preprint arXiv:1711.05101, 2017.
64. Y. Yan, Y. Mao, and B. Li, “SECOND: Sparsely embedded convolutional detection,” Sensors, vol. 18, no. 10, p. 3337, 2018.
65. C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, “Categorical depth distribution network for monocular 3D object detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8555–8564.
66. D. Park, R. Ambrus, V. Guizilini, J. Li, and A. Gaidon, “Is pseudo-LiDAR needed for monocular 3D object detection?” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3142–3152.