Multimodal Fusion Object Detection Technology in Intersection Traffic

Tianyuke Wang

{wangtianyuke@gmail.com}

School of Information Engineering, Chang'an University, Xi'an, Shaanxi, China

Abstract. The complex nature of the traffic scenarios in autonomous driving and the concept of the intelligent transportation system place high demands on perception system reliability. The sensors used by individuals are normally affected to a great extent in perceiving information when they are subjected to harsh environments such as harsh weather, lower light sources, and complex environments. The need to make use of the multimodal sensor technology has become urgent in creating overall visual and all inclusive perceptions. This paper takes a systematic analysis of the latest developments in this discipline, placing more emphasis on the analysis of three main issues, namely, the problem of feature complementarity by modal diversity, poor adaptivity to complex environments, and the failure to recognize long-tail targets and open environments. To overcome these hurdles, this paper goes further to expound on the common coping styles, propose commonly used datasets in this technology field and finally discusses the areas of concern, potential research opportunities and emerging trends with regard to multimodal fusion object detection technology.

Keywords: Autonomous driving; Intelligent transportation systems; Multimodal fusion; 3D object detection; open-world object detection.

1 Introduction

The strong incorporation of artificial intelligence, big data and Internet of Things technologies is making it easier to transfer autonomous driving and intelligent transport systems models to the phase of large-scale implementation with the involvement of a technical experiment. Those developments are on the rise as one of the key factors to promote the higher efficiency of traffic and the revolution of the travel system. Their development is closely related to the smart enhancement of road safety and road control. One of the core challenges of the perception layer is object detection, which is charged with the responsibility of detecting different participants of the traffic and other environmental components, hence forming the basis of future decisions and planning. The reliability of the system is also directly related to the accuracy, the real-time performance, and the strength of object detection [1]. The complexity of traffic situations, however, is full of interferences, which make detection of the target a major challenge. Individual sensors have intrinsic drawbacks: cameras are extremely sensitive to the light effect, lidar is expensive in terms of its computational costs and has

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365352

semantic challenges, and point clouds are readily susceptible to the weather [2]. Also, millimeter-wave radar does not have a high resolution, thus, it is very hard to locate small objects precisely [3]. These limitations lead to false alarms and false detections with single-modal detection, low flexibility to different scenes, and failure to meet the needs of a holistic perception of the scene.

Over the last few years, the development of deep learning, Transformer architecture, and cross-modal alignment technology has resulted in the development of the multi-modal fusion object detection technology as a key solution to dealing with the shortcomings of separate sensors and exploiting the complementary capabilities of sensors with different modalities. The technology is a form of technology that combines heterogeneous information from various sources, making use of the complementary information of the other modalities, to substitute the weakness of a single modality. It allows factoring in targets, localization and detection of the targets in complex surroundings accurately and this boosts the strength and stability of perception systems. Finally, this technology has been vital in making autonomous driving and intelligent transport systems a prevalent practice.

As a reaction to these challenges, the paper mechanically examines the progress in multimodal fusion detection of objects. It distills the history of and technical bottlenecks of current approaches in three modes: the modal heterogeneity of processing, robustness in complex environments, and generalization in long-tail scenarios and open scenarios. It is on this basis that the paper discusses the possible future research directions that should be carried out to act as a guide towards creating one of the most trusted perception systems that can be applied in complex traffic situations.

2 Issues and Progress in Modal Heterogeneity

One of the ways in which modal heterogeneity can be dealt with at the first stage is to map the point cloud onto the image and add to the image pixels a depth RGB channel (RGB-D). This is to ensure that the spatial features in the image are consistent with the spatial features of the point cloud features after fusion. However, in this case of feature integration, there is a possibility of major loss of spatial information [4].

To address this problem, other scientists have come up with the Lift-Splat-Shoot three-step paradigm. It is a depth prediction method of an image that embeds features of the image in 3D. This method greatly increases the depth of an image by predicting depth as a 3D distribution in an image instead of a 2D distribution. At the same time, the point cloud is voxelized and mapped to the Birds-Eye View (BEV) space to remove distortion of perspective and ensure the size of the target remains unchanged [5]. Later, the BEV space has become the focus of the activity of numerous researchers. To achieve that, BEVfusion builds an effective and symmetrical comprehensive framework of BEV representation, forming a common intermediate image and point cloud project space into the BEV space, respectively (Fig. 1) [6]. Such an explicit projection is very reliant on the position where the sensor is installed and the depth estimation performance. To handle these difficulties, Li Zhiqi proposed self-attention and cross-attention in both time and space, introducing BEVFormer [7], which makes use of a deformable attention model to help transform multi-camera images to the BEV space successfully. The technique obtains very high-accuracy centration of point cloud details.

Based on these benefits, Edmundo Guerra has included CBAM attention in the BEVfusion [8], allowing the model to handle key information when solving dynamic targets and dynamic edges by dynamics feature calibration both in channel and spatial scales.

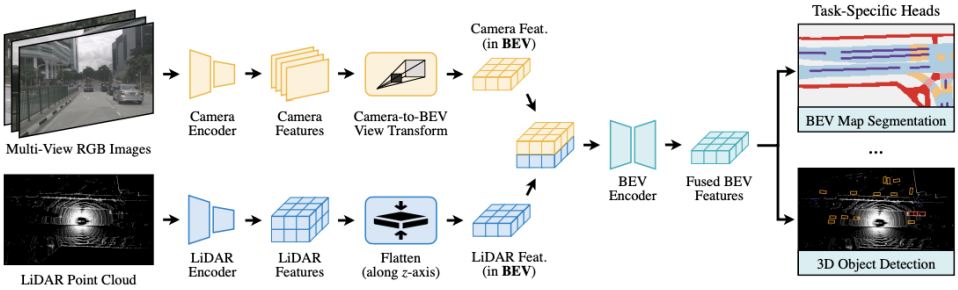

Fig. 1. Schematic diagram of BEVfusion [6]

The Multimodal large models have brought new ideas to the world of object detection. CLIP is a dual-encoder format that is composed of the image and text encoders to perform image-text feature alignment with a large-scale contrastive learning structure. Therefore, the third modality is natural language, which can be transferred using transfer learning [9]. This motivated the design of the pointCLIP method [10], which maps point clouds in different perspectives to depth images, thus introducing the point cloud features to a more similar level to the image features at the semantic level. The characteristics of both modalities are then mapped to a common semantic text space, to enable the consistent mapping and alignment of fine-grained semantic data of images with geometric structure data of point clouds at the natural language level. Such semantic consistency adds some semantics to the context in visual features, and it helps models understand picture content in a blurred or complicated image.

3 Problems of adapting and advancing in complicated settings

Complex environments have significant variation in sensor sensitivity to various modalities. In environments with a high intensity of light, sharp variations of light and darkness, and low light, visible light cameras can suffer problems with restricted dynamic range and distorted exposure. Such problems may cause reduced quality of imaging and destruction or inability of the target semantic information. In the case of LIDAR sensors, the raindrops, snowflakes and water stains may cause any kind of reflection or scattering to the emitted laser beams due to the environmental conditions. This results in a significant enhancement of point cloud noise, degradation of effective echo signals and an enhancement of multi-path interference. This leads to the deterioration of the radar detection range, target localization, and the quality of perception of the environment, as well as the stability and reliability of recognition features.

Yang proposed an edge vision model, EvGNN, which is an event-driven graph neural network model in cases where there is not enough visible light [11]. This model encompasses event flow in the form of a directed dynamic graph, which has single-hop nodes and infinite storage. EvGNN allows reducing the dependence on lighting conditions since it uses the power of

event cameras to effectively represent local scene data in a spatio-temporally decoupled search space. It can successfully do detection and tracking operations but is not sensitive to dynamic blurring and changes in illumination. To resolve difficult imaging conditions and also advance the accuracy and robustness of target localization, scholars have come up with a LiDAR-Camera fusion system called TransFusions [12]. This method uses a Transformer attention mechanism that defines a soft connection between LIDAR point clouds and image pixels. The model adapts effectively to changing perceptual tasks in dynamic environments by modeling spatial structures and contextual relations, and thus adapts to the optimal image features. Upon adverse weather like rain and snow compromising point clouds, leading to noise and distortion, image information has the potential to offer reliable structural support and thus, the provision of point cloud data is crucial in terms of its availability and legibility. Also, it can automatically eliminate redundant and abnormal noise to practically mitigate the risk of false targets and false detections. It is worth noting that the model can also enhance its recognition capacity and the precision of positioning the hard-to-find objects in the point clouds with the help of an image-based query initiation strategy.

Influences of complex environments do not just end with simple physical interruptions. Whereas space and time are easily aligned in a controlled laboratory-based scenario, the actual conditions of traffic would provide great challenges in alignment and calibration of multi-model data, given that high-speed traffic on rough highway beds may be considered in this case. When cars are driven under such circumstances, external sensors will have a slight offset that slowly builds up with the driving distances and causes spatial misalignment and feature differences across the different modalities of information, which, in turn, influences the quality of multi-model fusion.

Even though fusion approaches, such as BEV Fusion or BEVFormer, can effectively reduce the issue of modal heterogeneity by unifying the space of representations and being more tolerant to the formation of slight alignment deviations to some degree, they remain unable to address the underlying issue of feature mismatch as a result of the continuous change of external parameters and cumulative errors. Thus, researchers have done intensive investigations based on sensor registration and calibration procedures. Based on the underlying geometric structure, RGGNet uses the constraints of deep graph neural networks to learn the global geometric structure of point clouds, combines a temporal graph to correct online the offset of external parameters and runs effectively to reduce the error accumulation process, as well as to eliminate the spatial misalignment issue among multiple sensors [13]. It can learn the implicit tolerance in the error range as well as decrease calibration errors by using end-to-end world registration. Online Extrinsic Calibration: RGGNet is also provided with real-time and stable initial estimation of external parameters and through the dynamic compensation of the instantaneous pose change during jolts and vibrations during vehicle operation, it again reduces the range of the registration search in the model, and also, the convergence rate and accuracy of global geometric matching are enhanced [14]. By contrast, PETR enhances local perturbation reduction and long-term drift compensation in the time direction by aggregating features of cross-frame target positions, and the spatial time grouping (Fig. 2) [15]. It is a dynamic compensation mechanism that would be effective to compensate for the asynchronous acquiring delay due to non-matching sensor sampling rates, and mitigate motion blur and pose error in high frame rate situations. It is also coding information on 3D coordinate position in the image features to create 3D position-aware features which can be

tolerated by the targets. Therefore, this technique is used extensively in the 3D target detection and segmentation in multi-view, which has shown great success.

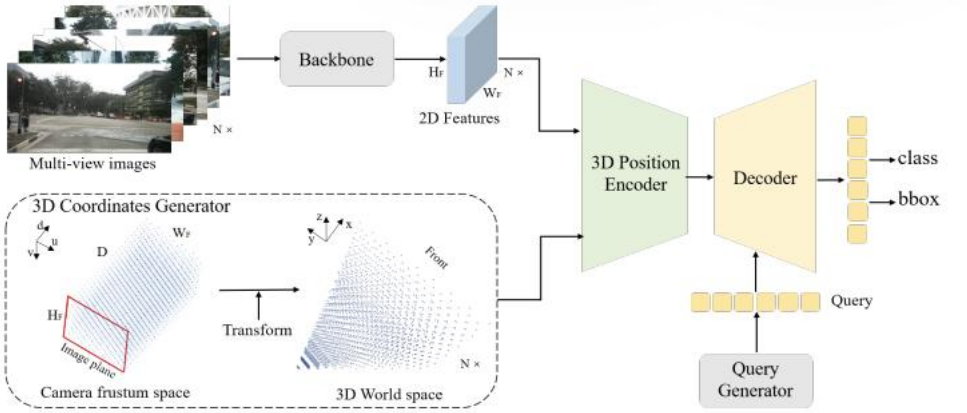

Fig. 2. Schematic diagram of the PETR paradigm [15]

4 Long-tail objectives and open-world developments and advancements

The categories of targets, which have extremely low occurrence probability and are of very small size, are called long-tail targets in complex traffic situations. The main issue is associated with unevenness of the sample categories: samples associated with top targets are readily available, and they contribute a significant part of the total sample, whereas tail samples are challenging to collect owing to the high costs of annotation. This is a disadvantage of posing a large quantity of high-quality data, obstructing sufficient learning of tail targets by the model and a resultant inadequate generalization capability. Due to this, the model tends to confuse the labels, features, and backgrounds of the highest targets and does not provide the desired detection performance. Furthermore, most of the current approaches do not utilize complementary information provided by multiple modalities to complement or improve the characteristics of tail targets; several optimization algorithms are also optimized only to evenly distributed targets, which, again, limits the training effectiveness of the model.

Currently, generative models such as GAN [16] and diffusion models [17] are capable of approximating the distribution of rare targets through data augmentation, which rectifies the lack of real samples, and training data are of high quality. Nevertheless, in difficult circumstances, the optimization of results in restoring and building characteristics of rare objects is difficult. Therefore, the model-level solution development has become a basic strategy to identify the infrequent targets in complex backgrounds. The improvement of the target loss function, optimization of the feature extraction network to improve the learning of rare features, and the use of an adaptive multimodal fusion approach can all be used to improve the detection accuracy of the model and its generalization of rarely occurring targets in extreme conditions.

The design of an adaptive loss approach, which gradually estimates the loss weights of the two targets, head and tail, increasing the loss proportion of the tail classes, and redirecting the model to discover the character of tail targets, is a critical plan to improve the performance of long-tail target detection. Focal Loss is an extension of the conventional cross-entropy loss, which adapts the weight of the loss of correctly classified samples in situ so that the model can focus more on the harder-to-classify samples and underrepresented categories during training [18]. Based on this idea, GHM Loss underlines the important role of the gradient regarding sample diversity. Using a gradient density normalization mechanism, it can help address gradient variations due to outlier gradients of extreme samples and increase the robustness of training of its model [19]. Taking into account the disproportionality of the values of the norms of the classifier weight, C2AM Loss does not use the operation of inner product, but resorts to maximizing the cosine similarity of the features obtained with the target category weight vector [20]. At the same time, it adds category-conscious adaptive Angle margin that shifts the decision boundary to a moderate extent of the categories with lower weight-norm values, hence maximizing spatial allocation and augmenting the recognition performance of tail categories.

On optimization of feature extraction networks, to mitigate the issue of the lack of feature learning due to the lack of long-tail samples, the existing studies focus on three primary directions, namely structural decoupling, knowledge transfer, and attention boosting. Forest R-CNN employs the relationship among target classes. Towards the goal of overcoming the logit noise and the category imbalance issues in the long-tail detection of a large number of target categories, it non-parametrically learns its hierarchical target instance study of parent category, and fine-grained category and constructs a classification forest tree-like hierarchical structure. Simultaneously, it uses a combination of the non-maximum resampling technique to even out the data distribution. The decoupling learning of the general features at the head and the exclusive features at the tail has been achieved, which overcomes the contradicting issue at the network structure level that a single backbone network will not be able to support different frequency targets [21]. To successfully transfer the knowledge on the sparse tail categories, SimLTD has adopted a three-step transfer learning approach where the head classes with rich samples are pre-trained, and then the knowledge is effectively transferred to the sparse tail categories. Limited sample conditions were found to be an efficient feature generalization when using unlabeled images, which do not need manual annotation [22]. The Long-Query Transformer extends query vectors and enhances the long-range attention generation. Through the assistance of denser query features and world context relationship functionalities, it also boosts the capacity of the model to learn and model long-tail targets with sparse distribution and weak features. Thus, in order to obtain a very accurate extraction and efficient depiction of the prominent characteristics of the tail target [23].

Regarding the multimodal fusion strategies, to continue to make use of the heterogeneous information to compensate for the drawbacks of single modal in long-tail target perception, current approaches are gradually advancing how to detect sparse categories in complex traffic scenarios with a continuous enhancement of basic model and semantic priors injection, etc. In MDETR, Aishwarya Kamath attains a rich mixture of text semantics and image characteristics, trains on large-scale text-image harmony data, and increases attribute distinction of fine-grained and rare targets by means of natural language understanding. It overcomes the bottleneck of a predetermined vocabulary at the semantic level and enhances

the discrimination capacity of the model with long-tail categories greatly [24]. ViLD applies the knowledge distillation technique on both the vision knowledge and the language knowledge that the pre-trained classification model has been trained to learn to distill the knowledge of the two-stage detector into the pre-existing embeddings of the detection box, replacing the text and image embeddings of the classification model. This is an effective expansion of the category space without the large-scale labeling, as well as a significant decrease in the use of rare category samples [25]. Being the most recent embodiment of the improvement of the fusion basic model, FOMO-3D is better adapted to intricate traffic conditions. It follows a two-step detection paradigm, applies abundant semantic and depth prior introduced by OWLv2[26] and Metric3Dv2[27] into the 3D perception framework and generates the image semantic data within OWL, dynamically improving the description of the sparse clouds of points. This allows it to continue with powerful feature support to long-tail targets in hostile conditions and severe traffic (Fig. 3) [28]. The basic techniques in the domain of open vocabulary detection and depth estimation are OWLv2 and Metric3Dv2. Since it is a space issue, they will not be discussed here.

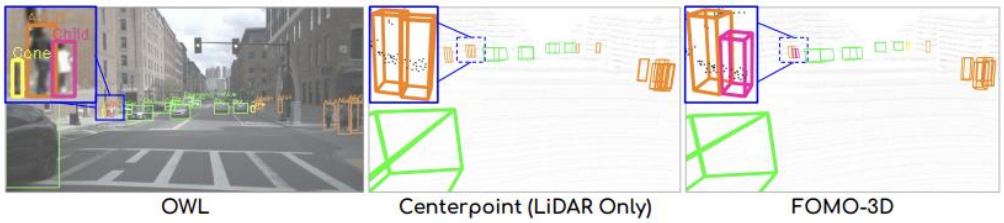

Fig. 3. Comparison of qualitative 3D detection results (OWL vs.Centerpoint vs FOMO-3D) [28]

In natural scenarios, in addition to the long tail targets that pose huge problems in detecting tasks, dynamic and uncertain target types pose significant complexities to perceiving targets in a complex traffic environment.

During training, the model is introduced to a limited number of familiar category targets, but unfamiliar category targets that are not trained might arise in actual situations. Furthermore, the fact of category distributions is associated with dynamic changes based on different regions and the environment. Open-world detection technology that builds on vision-language pre-training has become the most popular method in addressing these concerns due to its strong semantic alignment and zero-shot generalization features.

The combination of the object detection task and the phrase localization task is first achieved by GLIP. It achieves a visual representation model that is able to simultaneously learn object-level features, language perception information, and rich semantics through intense visual-language cross-modal interaction [29]. This list of detecting data and positioning data can be used to simultaneously learn the model using this unified modeling method, and therefore, improve the performance of the two activities simultaneously. Simultaneously, it also has the ability to create positioning boxes on large-scale graphic and text data automatically by means of self-training, such that more semantically important and more accurate features are learned comprehensively by the model. Youdaoplaceholder0 Dino, in its turn, and even more refines the unified framework of GLIP as it incorporates the Transformer-based detector DINO along with the depth of localization pre-training to build an open set detection model that is able to

detect any target via text categories or description instructions [30]. It follows a modality of language in a closed-set detection framework to achieve the generalization and broadening of the open set concept and splits the detection process into several essential steps. Small modules of fusion structure, like feature enhancers, feature selection with language guidance and cross-modal decoders, are developed. On the assumption of preserving the leading ideology of language, a new strategy of cross-modal feature combination of the whole stage and a more sophisticated query interaction mechanism has been provided. This has allowed the model to heavily integrate visual properties with text semantics, offering superior quality topic localization as well as free-world recognition with complicated text directions. OWL-ViTv2 begins with the intrinsic alignment property of the pre-trained model, constructs a simple and open vocabulary detection system that is based on CLIP, achieves the large-scale scale development of the detection data via the self-training mechanism, and applies the existing detector to automatically processor generate pseudo-box annotations on the large-scale image-text pairs to further boost the alignment effect of the vision and language [26]. It is important in edge detection activities where the computing power resources are limited because it supports a lightweight architecture and is very cheap in self-training, and its performance in real-time is very good. In the meantime, a new end-to-end open-world detection technique, OW-DETR, based on deformable attention and open set query design, has an excellent global multi-scale context capturing capacity and vastly powerful ability to transfer knowledge, and thus, further enhances the architecture system of open-world detection in complicated traffic settings (Fig. 4) [31].

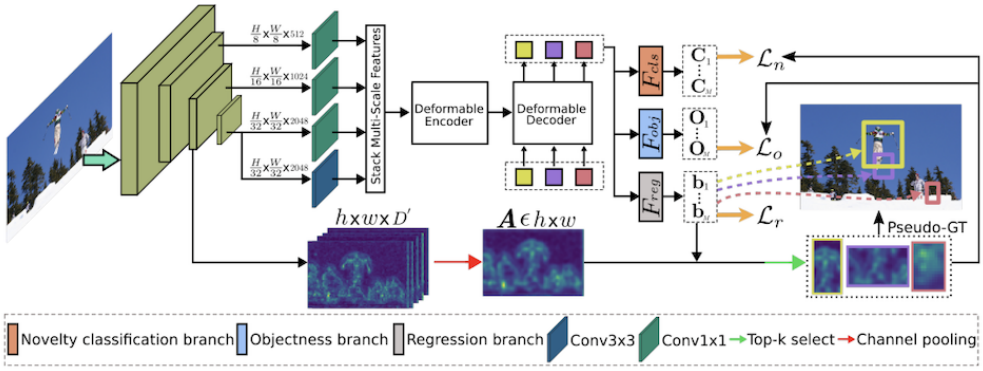

Fig. 4. Schematic diagram of the OW-DETR framework for open-world object detection [31]

5. Dataset

The most popular datasets in the sphere now are the KITTI, the nuScenes, the Waymo, and the V2X-Radar.

KITTI is a database that provides 389 image stereo and optical-flow, a sequence of visual odometry over a distance of 39.2 kilometers, and over 200, 000 images that are annotated with 3D objects. This dataset offers a wide range and quality assessment data in order to implement different computer vision activities, such as stereo vision, optical flow estimation, visual odometry, 3D object identification and multi-object tracking.

The nuScenes dataset consists of 1, 000 natural traffic scenes in Boston and Singapore, with different conditions, such as congested traffic, rainfalls, dark nights, and cars with steering wheels on both sides. In every scene, the average duration of each scene is 20 seconds, and it is marked with 3D bounding boxes having eight attributes and 23 categories. Because of its elaborate sensor configuration, a wide variety of complicated scenes and its refined 3D markings, nuScenes is licensed to autonomous driving studies to help in testing algorithms, as well as monitor effectiveness in enhanced undertakings, encompassing 3D object recognition, pursuit tracking, point array semantic mapping, and birds-eye perception (BEV).

The Waymo data set consists of 1, 150 real-life situations mostly located in the highly populated urban systems, suburbs, business districts, and residential neighborhoods. It is superior in demonstrating diverse difficult scenarios such as problematic traffic, extreme barriers, and small roads, as well as complicated crossways, and emergency cases, overcoming the intricacy of similar sets of data. It is important to note that the Waymo dataset is a rich resource of high-precision 3D annotations, with a large concentration of LiDAR point clouds, exhaustive annotation of targets, and outstanding positional accuracy. Notably, its richness is 15 times greater than the most available camera + LiDAR dataset, making it a rigorous standard of measurement in the analysis of the target detection and multimodal fusion perception under a complex traffic environment.

The V2X-RADAR dataset consists of 20, 000 LiDAR image frames, 40, 000 camera image frames, and 20, 000 4D radar data frames, each consisting of 350, 000 3D bounding boxes within 5 major scenarios, i.e., vehicles, pedestrians, and traffic cones. It deals with complex traffic scenarios that are highly occlusive and with strong interactions during diverse weather conditions and periods. The first 4D Radar fusion dataset, V2X-RADAR, can be subdivided into V2X-RADAR-C (V2X perception), V2X-RADAR-i (roadside independent perception) and V2X-RADAR-V (single-vehicle independent perception). The data is very useful in expanding the scope of the research on multimodal fusion and vehicle-road cooperative perception.

6. Limitations of the Present and Future

6.1 Existing Limitations

The multimodal fusion technology, which is critical towards combining dissimilar data, includes complex operations such as feature extraction, cross-modal alignment, and fusion inferences. Although modal alignment accuracy can also be improved when the Bird Eye View (BEV) space is considered, feature projection coupled with spatial transformation is a significant burden in the realization of the model with huge proportions that slow down inference rate. Satisfying real-time requirements of target detection in an autonomous vehicle cannot be compromised without a powerful computer.

Although the above-mentioned optimization of the loss function, enrichment of feature extraction and application of semantic priors can alleviate the long-tail target detection problem, the schemes are often specific to particular complex situations and, therefore, are not very general and are not able to generalize their performance when there are changes in the target distribution in different settings.

In non-directional detection, existing methods tend to rely a lot on large gatherings of image-text matching, and also pre-trained models, and may result in feature solidification of the models. This may result in a significant reduction in recognition accuracy to unfamiliar target categories, as well as impair the dynamics of target category distributions of novelty and complex scene properties in the open world problem. In addition, they usually do not offer great semantic reasoning capacity in relation to the behavior of unknown objects, thus not sufficiently serving the needs of the perception of targets and decision-making tasks in the real-world autonomous driving conditions.

6.2 Future outlook

The future of multimodal fusion target detection technology will be advanced to high-efficiency, real-time, low-mass, high-intelligence, based on the special conditions of complex traffic and the innovations in further sensor technologies, as well as the further improvement of artificial intelligence.

One, the use of more efficient and light converged architecture can be useful to reduce the scale of parameters in the model and the complexity of its computation and maximize its inference speed. Moreover, edge computing research is suggested to be combined with multimodal objects detection technology that may allow putting all the advantages of edge computing processing power to the forefront, reducing the daily losses of data transportation, and further increasing the efficiency of fusion.

Moreover, to improve long-tail target detection and open-world detection, the focus should be on taking advantage of multimodal semantic information and working out the solutions to the few-shot and zero-shot learning. Improve generative data augmentation methods to produce more realistic target samples that reflect a more realistic traffic situation in the real world. Also, further enhance the merging of the vision-language prepared models and multimodality to improve the recognition achievement of the model on unknown target categories. Further, the model has the capability of swiftly absorbing the features of new target groups when combining practices like transfer learning, reinforcement learning, and incremental learning, enabling the model to detect and track dynamic data.

Multimodal object detection technology will continue to be developed in the future, together with other advanced types of technologies such as vehicle-road cooperative perception, large language model reasoning, in an attempt to come up with a more complex and efficient visual perception system. Vehicle-road cooperative perception combines perception data of different sources(vehicles, roadside devices, and cloud platforms) to attain global perception and collaborative decision-making. In the meantime, Large language models are strong based on the semantic understanding and reasoning, support semantic alignment and targeting behavior reasoning of the multimodal data, thus enhancing the researchers to strengthen the model's performance.

7. Conclusion

The object detection technology is also an emerging research area in the field of autonomous driving and intelligent transportation, targeting promising research prospects. The paper covers the developments in multimodal fusion object detection technology in complex traffic

scenarios, highlights the main problems that such technology faces, outlines the main challenges of each problem as well as some prominent solutions to these problems, and supplements the major sets of data they used in these research directions. It eventually summarizes the existing limitations and direction of multimodal fusion object detection technology in multifaceted traffic scenes in the future.

References

[1] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 580-587 (2014)

[2] Zhou, Y., Tuzel, O.: VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4490-4499 (2018)

[3] Rohling, H.: Radar CFAR thresholding in clutter and multiple target situations. IEEE Transactions on Aerospace and Electronic Systems. 19(4), pp. 608-621 (1983)

[4] Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. International Journal of Robotics Research. 31(5), pp. 647-663 (2012)

[5] Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In: Proceedings of the European Conference on Computer Vision. pp. 424-441 (2020)

[6] Liu, Z.J., Tang, H.T., Amini, A., et al.: BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. arXiv preprint arXiv:2205.13542 (2024)

[7] Li, Z.Q., Wang, W.H., Li, H.Y., et al.: BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. arXiv preprint arXiv:2203.17270 (2022)

[8] Zhang, N., Guerra, E., Grau, A.: AttBEV: Enhancing Multi-Modal 3D Object Detection With CBAM Attention in BEVFusion for Autonomous Driving. IEEE Robotics and Automation Letters. 11(2), pp. 1322-1329 (2025)

[9] Radford, A., Kim, J.W., Hallacy, C., et al.: Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020 (2021)

[10] Zhang, R.R., Guo, Z.Y., Zhang, W., et al.: PointCLIP: Point Cloud Understanding by CLIP. arXiv preprint arXiv:2112.02413 (2021)

[11] Yang, Y.F., Kneip, A., Frenkel, C.: EvGNN: An Event-Driven Graph Neural Network Accelerator for Edge Vision. IEEE Transactions on Circuits and Systems for Artificial Intelligence. 2(1), pp. 37-50 (2025)

[12] Zhou, C.T., Yu, L.L., Babu, A., et al.: Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. arXiv preprint arXiv:2408.11039 (2024)

[13] Yuan, K.W., Guo, Z.Y., Wang, Z.J.: RGGNet: Tolerance Aware LiDAR-Camera Online Calibration With Geometric Deep Learning and Generative Model. IEEE Robotics and Automation Letters. 5(4), pp. 6956-6963 (2020)

[14] Cocheteux, M., Moreau, J., Davoine, F.: Uncertainty-Aware Online Extrinsic Calibration: A Conformal Prediction Approach. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6167-6176 (2025)

[15] Liu, Y.F., Wang, T.C., Zhang, X.Y., et al.: PETR: Position Embedding Transformation for Multi-View 3D Object Detection. arXiv preprint arXiv:2203.05625 (2022)
[16] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., et al.: Generative Adversarial Networks. arXiv preprint arXiv:1406.2661 (2014)
[17] Yang, L., Zhang, Z.L., Song, Y., et al.: Diffusion Models: A Comprehensive Survey of Methods and Applications. arXiv preprint arXiv:2209.00796 (2025)
[18] Lin, T.Y., Goyal, P., Girshick, R., et al.: Focal Loss for Dense Object Detection. arXiv preprint arXiv:1708.02002 (2018)
[19] Li, B.Y., Liu, Y., Wang, X.G.: Gradient Harmonized Single-stage Detector. arXiv preprint arXiv:1811.05181 (2018)
[20] Wang, T., Zhu, Y.S., Chen, Y.Y., et al.: C2AM Loss: Chasing a Better Decision Boundary for Long-Tail Object Detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6970-6979 (2022)
[21] Wu, J.L., Song, L.C., Wang, T.C., et al.: Forest R-CNN: Large-Vocabulary Long-Tailed Object Detection and Instance Segmentation. arXiv preprint arXiv:2008.05676 (2021)
[22] Tran, P.V.: SimLTD: Simple Supervised and Semi-Supervised Long-Tailed Object Detection. arXiv preprint arXiv:2412.20047 (2025)
[23] Askari, A., Verberne, S., Abolghasemi, A., et al.: Retrieval for Extremely Long Queries and Documents with RPRS: a Highly Efficient and Effective Transformer-based Re-Ranker. arXiv preprint arXiv:2303.01200 (2023)
[24] Kamath, A., Singh, M., LeCun, Y., et al.: MDETR -- Modulated Detection for End-to-End Multi-Modal Understanding. arXiv preprint arXiv:2104.12763 (2021)
[25] Gu, X.Y., Lin, T.Y., Kuo, W.C., et al.: Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. arXiv preprint arXiv:2104.13921 (2022)
[26] Minderer, M., Gritsenko, A., Houlsby, N.: Scaling Open-Vocabulary Object Detection. arXiv preprint arXiv:2306.09683 (2024)
[27] Hu, M., Yin, W., Zhang, C., et al.: Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-Shot Metric Depth and Surface Normal Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 46(12), pp. 10579-10596 (2024)
[28] Yang, A.J., Tu, J., Dvornik, N., et al.: FOMO-3D: Using Vision Foundation Models for Long-Tailed 3D Object Detection. In: 9th Annual Conference on Robot Learning (2025)
[29] Li, L.H., Zhang, P.C., Zhang, H.T., et al.: Grounded Language-Image Pre-training. arXiv preprint arXiv:2112.03857 (2022)
[30] Liu, S.L., Zeng, Z.Y., Ren, T.H., et al.: Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv preprint arXiv:2303.05499 (2024)
[31] Gupta, A., Narayan, S., Joseph, K.J., et al.: OW-DETR: Open-world Detection Transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9225-9234 (2022).