Algorithms and Frameworks for Multi-Sensor Collaborative Perception in Autonomous Driving

Feiyang Lin

{136152023096@student.fjnu.edu.cn}

College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou, 350117, China

Abstract. The recent years have seen work done on algorithms and frameworks of multi-sensor cooperative perception. This is an emphasis meant to improve the environmental perception of the autonomous vehicles. The heterogeneity of the multimodal data can be viewed as one of the major challenges during the multi-sensor fusion process. Another major challenge is in difference in the fusion levels. Also, model computational efficiency is also an important research problem. The paper will begin by giving the nature of popular sensors. It goes on to describe pertinent research areas in multi-sensor fusion in autonomous driving. Advanced paradigms of fusion symbolized by BEV and Transformer are analyzed in detail. New vehicle-infrastructure vehicles co-operative fusion structures are discussed too. Lastly, the paper will conclude by summarizing the existing issues and predicting the future trends. A perceived and decision making and planning is one of the possible development directions in the future.

Keywords: Autonomous driving; Multi-sensor fusion; Collaborative perception algorithms

1 Introduction

The current stage of development of autonomous driving technology is quite rapid. Safe and stable functioning of autonomous vehicles is strongly related to the ability of the system to retrieve environmental data. Its efficiency in processing is also a determining factor that is critical [1].

The conventional methods of perception are often based on the use of one sensor to collect data. There are several limitations that are usually related to this approach. The single sensor perception performance is inadequate in tricky conditions of roads. It cannot achieve the high-accuracy and low-latency nature of autonomous driving. As a result of this, the multi-Sensor fusion paradigm has surfaced. Multi-sensor fusion can be defined as the process of collecting data with the help of multiple sensors, i.e. LiDAR, cameras, and millimeter-wave radar. It satisfies the objective of the precision of joint perception of the environment with the integration of data on the basis of corresponding algorithms. In this integrated scheme of perception, there is more opportunity to capitalize upon the various peculiarities of sensors in regard to the

ICMEEA 2026, April 24-26, Singapore, Singapore

DOI 10.4108/eai.24-4-2026.2364940

resolution and detection range. This approach enables it to attain the complementary advantages [2]. Researchers have found several directions of multi-sensor fusion implementation to autonomous driving in recent years. Each of these levels can be divided into three major levels with regard to fusion strategies: there are data-level, decision-level, and feature-level fusion strategies. Among them, feature-level fusion, owing to its trade-off between retention of information and learnability, has been one of the mainstream research focuses. On algorithm and framework level, the focus efforts are mainly allocated towards the construction of efficient and iterative platforms of analysis. BEVFusion framework is a combination of the spatial features of multi-modals by use of the Bird's Eye View (BEV) representation, thus, being successful in capturing more of the spatial information [3]. V2I-MSF framework is a framework that considers a new opportunity in connecting roadside infrastructure with vehicle-mounted sensors, which proves to be effective in urban complex settings [4]. Study findings show that multi- sensor fusion has strengths of improving environmental perception of autonomous driving. Nevertheless, unresolved vices are still present in this technology itself. The current paper dwells upon LiDAR and camera-based multi-sensor fusion algorithms and frameworks. It surveys the evolution of technology since the principles of sensors, to state of art perception algorithms. The literature also goes further to the study development of car-infrastructure collaborative perception. The paper also examines what is the current issue with the multi-sensor fusion technology and what directions the technology may take in the future.

2 Sensor principles and characteristics

The autonomous driving is one of the areas that sensors are used universally in two types. The former includes vision cameras to detect images including monocular and stereo cameras. The second one incorporates radar-based sensors, including LiDAR and millimeter-wave radar. The two types of sensors noted above have major differences in terms of how they operate as well as advantages and disadvantages associated with their operation.

2.1 Monocular camera

A monocular camera is an optical sensor which employs one camera to record the data in the form of an image of the surrounding in a single imaging view. Such sensor widely works on the pinhole imaging model. It trains the spatial characteristics of a three-dimensional setting to a two- dimensional imaging plane. The processing results in the derivation of an RGB image of the captured region. The monocular camera has a restricted resolution to its imaging concept and therefore, its visual acuity on the environment is based on sensor pixel outcomes. It is also dependent on the processing power of latter algorithms. The monocular camera image data lacks actual depth data of the surrounding. This therefore makes it hard to have the accurate judgments in relation to the distance of the target by just using such image information.

2.2 Stereo camera

A system that consists of two cameras which are used to view the world around in a similar way as how the human eye sees through binoculars. It does the 3D data processing of the environment through the use of parallax. The depth information of the target with respect to the stereo system can then be determined by comparing the difference of lateral position of the same

target as against the background of both cameras, and together with triangulation algorithms. This makes the convenient means of measuring distance fairly accurate.

2.3 LiDAR

LiDAR senses an accurate distance of the targets by calculating the aggregate elapsed time of the emitted laser pulses and the reflected significant reverted pulses. LiDAR is less sensitive to the lighting conditions and target color, as compared to vision sensors. Its performance is, however, vulnerable to the materials of the object surface and to the hiding behind situation by obstacles. LiDAR is actively used in the area of multi-Sensor fusion, in particular, in autonomous driving because it has rather high accuracy of the range and high quality of work. It is often used with cameras in order to have complementary benefits among various sensors. Table 1 describes the main features, pros and cons of cameras and LiDAR.

Table 1. Comparison of key characteristics between camera and LiDAR.

Type	Camera	LiDAR
Perception range	150-200 meters	150-500 meters
Data types	2D RGB/grayscale image	3D point cloud
Advantages	Rich texture dtails	High-precision ranging
Disadvantages	Susceptible to lighting and weather conditions	Relatively high cost

3 Levels of multi-sensor fusion

Multi-sensor fusion levels are classified into several ways. According to the type of data, it can be divided into data-level, feature-level and decision-level fusion. Data-level fusion is in which the sensors do not process data following collection. It incorporates direct integration of the raw data. This integrated data is then further subjected to extraction of features. Decision-level fusion on the other hand asks one to initially process the information of each sensor individually using a sequence of steps. The results of each single sensor are then obtained and fusion is then done based on the individual sensor independent results. The combination of this produces the end decision. The feature-level fusion deals with the feature extraction being performed on the raw data acquired by individual sensors and then the features obtained are individually provided. These feature sets which are extracted are then aligned and integrated. Fused feature information is further converted to be the basis of decision-making [5].

A comparatively low level in multi-sensor fusion is data-level fusion. The labor performed at this level is often directed at the purpose of attaining the correspondence and interoperability between the information obtained in various sensors. One popular method consists of converting LiDAR point cloud data into other formats, e.g. BEV images. This conversion allows combinations with RGB images of cameras. Alternatives, images on cameras are translated into virtual point clouds. LiDAR point cloud data is then combined with these virtual point clouds. Multi-sensor fusion through feature-level fusion is one of the frequently employed methodologies, with its major classes being feature concatenation, weighting algorithms, and attention mechanisms [6]. Over the last few years, the fast growth and advancement of the deep

learning technology have caused the further functioning of the similar feature fusion algorithms to be iterated. Nowadays, conventional attention mechanisms are capable of gaining dynamic changes of weight values thus increasing the accuracy of the capabilities of perception in different environments. The decision-level fusion mainly uses the approach of the voting method, Bayesian fusion and meta-classifiers. There is the meta-classifier that is one of the methods frequently applied in decision-level fusion. It uses the output of sub-models as its input and retrains another fusion model, which has a level of adaptive ability.

Various levels of fusion undergo the differences with regard to more than one thing. Their features, strengths, and weaknesses are listed in a summary in Table 2.

Table 2. Comparison of data-level, feature-level, and decision-level fusion.

Level	Data-level fusion	Feature-level fusion	Decision-level fusion
Fusion stage	Raw data stage	Feature representation stage	Decision stage
Fusion object	Raw sensor data	Feature information	Independent predictions of each sensor
Advantages	Theoretically minimal information loss	Suitable for deep learning	Good fault tolerance
Disadvantages	High complexity	Requires high-precision data alignment	Substantial information loss

4 Deep learning-based multi-sensor fusion algorithms

4.1 Fusion methods based on unified representation

The features of various sensors in the multi-sensor fusion process are usually data of varying type and temporally asynchronous, and there is some degree at which the features in the fusion process are misaligned. In primitive autonomous driving perception systems, the confusion problem of cross-modal information had not been fully addressed. Consequently, a multi-sensor fusion scheme has been suggested by the researchers on the basis of Bird's eye View (BEV) as a common space of representation. This architecture is expected to accomplish the combination of camera and LiDAR in the same spatial space. It aims at improving the correspondence between various sets of information on features.

In 2023 Liu et al. came up with a BEV model named BEVFusion, which was designed to do multi-sensor fusion [3]. It is one of the most outstanding representatives of recent years who uses the BEV as a single form of fusion. BEVFusion, as they integrate the feature representation of both cameras and LiDAR with the BEV and develop a general multi-task multi-modal fusion framework on this basis, attains some improvements in feature representation and fusion. In BEVFusion, feature encoders of various sensors are designed. These encoders obtain useful feature data and achieve the conversion to the BEV. With this, the semantic and geometric information of the environment can be maintained at the same time. Researchers have established that the pooling performance of BEV is relatively not good when converting the information about camera features to BEV representation because the size of camera feature point clouds are large. To overcome this issue, the framework makes the BEV pooling highly efficient by pre-classifying the mapping relationship and scope of computation and consequently causing the cost of computation to decrease. Once the BEV-based unified

representation has been obtained, one is able to fuse features of various modalities with element-wise operations. The BEV images produced by the LiDAR and the cameras can also be misaligned in their local areas due to camera depth estimation errors. Due to this reason, BEVFusion presents a convolution-type BEV encoder and an integration of residual forms to modify the fused features to enhance the alignment of multi-modality features.

It is still going through repeated iterations of perception frameworks based on the use of BEV as a single representation. Along with the BEVFusion framework, the LSS framework suggests a broad approach to extract the multi-view camera features on the BEV, which prepares a significant cornerstone on the design of the visual branch of the later multimodal BEV fusion models. LSS is a type of BEV framework that is a camera-based one. To improve the poor fusion rates of the depth information uncertainty of the camera image pixels, the LSS architecture projects multi-view image characteristics to the BEV using three-dimensional projection to provide an end-to-end transformation of multiple camera images into BEV representations [7]. Each camera image is first processed with LSS. It picks the image features along the direction of view to various discrete positions of depth. It shapes frustum representation containing semantic information. The frustum features are then brought together and combined into a discretized BEV grid. It is based on this that the LSS framework goes further to transform the features of the BEV into a cost map. It then superimposes pre-determined trajectories of templates onto the map. This does realize end-to-end motion trajectory planning. Being a traditional BEV-based framework, the LSS offered the “Lift-Splat” approach to viewpoint transformations has offered a reference point to the subsequent design of BEV perception framework.

BEVFusion has better opportunities in terms of the variety of feature sources and emphasizes more on cross-modal learning of multi-sensor features in comparison with LSS framework where the camera images are employed as the medium. The BEVFusion model takes advantage of the complementary nature of sensor features data. It is then able to more effectively hold the geometric information of LiDAR and the high-density semantic information of cameras. It also saves time that would be spent in turning the camera pictures to BEV. BEVFusion is less expensive to compute and has a higher pace in comparison with the past solutions. It is an effective perception structure [2, 3].

Other unified representations like front view and voxels are also considered to be of importance in autonomous driving environment perception besides BEV. They each have their sceneries to apply. The front view has a primary meaning of a representation that utilizes front-facing cameras. It recognizes and displays the surrounding in the forward-facing view of the vehicle. The front view being a visual perception solution, is commonly adopted as a carrier space that involves semantic knowledge of the environment in the multi-sensor fusion case. It does this to enhance multimodal data fusion by projecting data on sensors like LiDAR to the image plane. Nevertheless, due to the impossibility of cameras to capture the actual depth and the frequent inconvenience of experiencing an obstruction, there are some limitations to the front view [8].

The three-dimensional unified representation called voxels are also common. LIDAR point cloud data are commonly represented by them. In 2024, An introduced the TransVoxelRadar system which employs a structured voxel grid to store point cloud data, and combines local and global data around a voxel of view to obtain a structured spatial result [9]. There is good spatial consistency in voxel representation. Using voxels as unified representation can easily handle

three-dimensional spatial information and has found application in 3D perception. But as the size of data becomes large, as the voxel resolution rises, the computational cost of the model correspondingly rises dramatically. Due to this reason, the design of perception frameworks can be used sparsely voxel and such like algorithms to consider a balance between the computational cost and accuracy.

4.2 Transformer-based fusion methods using neural network architectures

Transformer was initially applied to natural language processing. Nevertheless, its strengths in terms of capturing long-range dependencies and modeling are in line with the autonomous driving multi sensors data fusion requirements. Transformer architecture has slowly become popular and recognized by researchers in the past several years as the way to process multimodal heterogeneous data [10]. Transformer architecture is a deep learning model that is built upon the attention mechanism. The attention mechanism supported by the Transformer architecture considers the information on the various modalities as being represented by a sequence of features during the heterogeneous data fusion process. It determines how each of the elements in input feature sequence is correlated with other elements to assign dynamically fitting weights [11]. This implies that the attention mechanism is able to circumvent single-step feature data superposition. Each of the features has an ability not only to model the global dependencies but also to capture the local information.

In multi-sensor fusion tasks, the data in the sensor modalities are first transformed to the same space of unified vectors by using the respective feature extraction networks. These are then inputted into the following Transformer computations in a feature map units transform. In the Transformer architecture, the attention mechanism provides a mechanism of associating the fusion weights of modalities to the query vectors of the decoder that perform different choices to incorporate the most pertinent feature information to the current query so that cross-modes could be interacted and fused. Through this mechanism, the model can be able to mediate information across various modalities in terms of meaning and time to present more comprehensive information about the environment.

Most frameworks that in the field of autonomous driving have been developed on the basis of the Transformer. Bai et al. created the TransFusion framework in 2022 that attained productive integration of LiDAR and cameras in low-image circumstances [12]. A convolutional backbone network and a detection head with the use of Transformers are components of the TransFusion framework. The TransFusion framework is able to select useful information in images and combine queries in LiDAR with image features using a two-layer decoder design and attention mechanism. Chen et al. created the FUTR3D architecture in 2023 [13]. The framework is a sensor fusion wide-angle 3D detection framework. It can be used on virtually any combination of sensors. The Transformer is also brought up by the FUTR3D framework. It also assists the algorithm in extracting features in each modality without depending on a certain modality, which enables the model not to be dependent on a particular modality. On the whole, the Transformer architecture can dynamically change the fusion weights of various sensor modalities based on them changing the input features. It is capable of controlling the input of both modalities to the fusion outputs. This enhances the strength of multi-sensor fusion processes and offers positional ability in terms of feature fusion.

5 Vehicle-infrastructure cooperative fusion

The above reviewed frameworks are largely focused on the autonomous vehicle. But in practice, the physical world causes the field of view of autonomous car to be easily obstructed by the presence of other buildings and other vehicles, clearing up to a low range. In order to alleviate this condition, vehicle-to-infrastructure (V2I) cooperative fusion schemes have been proposed by some researchers. Such plans are designed to allow multi-sensors to collaborate between roadside infrastructure and cars. V2I is a perception system that supports two-way capacity of information flow between vehicles and infrastructure along a road. To some degree, its appearance has increased the range of perception of autonomous cars to the level of the road. Roadside sensors are some of the significant elements of the system in this framework. Roadside sensors are also typically positioned with a greater height to gather data on a somewhat top-down point of view. This is useful in complementing the environmental data that the vehicles lose in blind spots of their field of view.

In 2025, the V2I-MSF framework was created by Khan et al. that combines onboard LiDAR and camera data with data obtained by sensors at the roadside infrastructure. This brings additional enhancement to precision and the strength of environment perception at challenging crossroads in the urban realm [4]. V2I-MSF perception system follows the principle of firstly merging the sensor data collected by the vehicle itself, and the data of sensors on the roadside. Images obtained by the cameras on the vehicle side majorly serve to get the semantic information of the surrounding. The algorithms used to detect objects include object detection algorithms like YOLOv4. Meanwhile, the LiDAR 3D points cloud data also computes the structural information and distance features. The convolutional neural networks are used to combine the data of the two sensors. Once fusion of the sensor data on the vehicle side is done, then the V2I-MSF framework transmits the fused data to perception modules on both sides of the road. Information merged on the car platform is further enhanced by cameras and LiDAR placed by the shoulder of the road creating a subsequent fusion of multi-data. The road side sensors are provided on a higher level to overcome the blind spots of the vehicle sensors. They acquire environmental information within a relatively broad viewpoint and supplement the information that is lost by the vehicle through occlusion. Besides this, to guarantee the viability and promptness of the fusion of multi-sensors between vehicles and roadside infrastructure, the V2I-MSF structure applies interpolation and resampling techniques to match the data. It also implements a delay-conscientious weighting approach in order to mitigate the effect of communication latency on the fusion results.

It currently ranks among the primary lines of autonomous driving cooperative perception, although numerous issues remain in the field of implementation. Under the V2I framework, the roadside sensors send their data to vehicles, using wireless communication networks. Communication latency and signal interference easily affect it during transmission and affect the real-time performance of cooperative perception. Moreover, V2I data are bi-directional and thus the continuous transmission of data for the two-way keeps on generating large amounts of data. The high mass of computations demands great exaltations on processing efficiency and communication bandwidth.

6 Current challenges and future outlook

6.1 Algorithm-level challenges

Most of the algorithms applicable to the realization of multi-sensor fusion in the field of the autonomous driving are currently rooted in the deep learning method. This implies that perception models use high-quality annotated datasets that are of large size. It is expensive to construct entire data sets of various urban conditions and weather situations. Given the variability of the surrounding when a vehicle is on the road, the use of training datasets makes deep learning algorithms in autonomous driving less generalizable and less transferable. The further work can be oriented on creating the networks that involve unsupervised or self-supervised learning [14]. In addition, real-world conditions involving high-risk events that are only covered by a few datasets are infrequent. They are low-frequency unusual situations that typically affect the system safety and trustworthiness [15]. Thus, the additional enhancement of the model in terms of his generalization capacity in long-tailed or extreme conditions is required. Conversely, in order to enhance the accuracy and strength of the multi-sensor fusion in condensing the environment perception, the existing multi-sensor fusion models are in general big parameter size and computationally intricate. This makes computation and storage of the perception system expensive. The way to realize model lightweighting and still retain the perception performance is now one of the concerns of multi-sensor fusion algorithms.

6.2 System-level challenges

Adoption of vehicle-to-vehicle (V2V) and V2I cooperative perception systems in vehicular communication systems (V2X) is one of the developmental trends in autonomous driving systems in the future. V2V and V2I solutions would adequately address the weakness of single onboard-based perceptions in comparison to the current solutions of cooperative perception [16]. Nevertheless, it is not easy to come up with a very feasible V2X system. In the case of V2V, the more vehicles involved in cooperative fusion, the worse the current wireless communication technologies cannot guarantee the legitimacy of communication among moving cars. Problems like packet loss as well as communication delay must be solved. In the case of V2I, sensors, performance of perception, communication latency, and spacing control strategies have differences when vehicles varying in models or brands use the same V2I system [17]. The safety of autonomous vehicles is somewhat compromised because of compatibility issues. Thus, enhancing compatibility of the V2I systems and establishing a uniform perception system are relevant in the future. In addition, when adapting the V2X systems to other vehicles, the issue of creating vast streams of computational data and real-time communication needs is a load to the communication reliability of the system. Currently, one of the challenges of developing a universal V2X system is still the high cost of research and development.

6.3 Future trends and outlook

Large-scale annotated data are used to build perception models and little generalization can be made in extreme cases. In order to overcome these shortcomings, scholars are trying to enhance simulation systems used in autonomous driving. Over the recent years, novel simulation technologies related to the emerging representation, including neural radiance fields (NeRF), have been evolving rapidly. The new simulation techniques could offer quality rendered simulation training data to autonomous driving perception models which would potentially

solve the challenge of inadequate and expensive datasets in the context [18]. They are also good in enhancing generalization capability of autonomous driving perception models in extreme conditions. Moreover, novel neural architecture search (NAS)-based frameworks are also developing quickly. With NAS new design capable of replacing manual neural networks, it becomes possible to obtain lightweight autonomous driving perception models, which will support the commercial production of autonomous vehicles on the scale of large size.

7 Conclusion

This paper starts by examining the development of multi-sensor fusion technology in autonomous driving environment perception. It systematically reviews the evolution from differences in sensor physical characteristics to the differentiation of fusion levels. It then covers cutting-edge frameworks based on deep learning algorithms and vehicle–infrastructure cooperative perception frameworks, tracing the main technological development trajectory. Studies have shown that the inherent complementarity of sensors such as cameras and LiDAR, in terms of data types and applicable scenarios, forms the basis for the scientific value of multi-sensor fusion technology. The evolution from data-level to feature-level and decision-level fusion reflects a trend from simple data superposition toward unified modeling and collaborative perception. With the help of deep learning techniques, various multi-sensor cooperative perception frameworks have flourished. BEV-based perception frameworks effectively address the alignment of multimodal data, providing a unified computational platform for multi-sensor feature fusion. At the same time, the Transformer architecture, with its attention mechanism, has been widely applied in autonomous driving multi-sensor fusion. It has become an important technique for handling long-range dependencies and achieving global modeling. The core of multi-sensor fusion lies in efficiently and robustly aligning and complementing heterogeneous data. BEV and Transformer are currently among the most promising technical approaches.

Multi-sensor fusion frameworks are still continuously evolving. In the future, V2I and V2V will be among the main directions for the development of autonomous driving perception systems. Future autonomous driving perception systems will no longer rely on the isolated intelligence of a single vehicle, but will be integrated into intelligent road networks capable of real-time cooperation. They will enable beyond-line-of-sight information exchange, joint perception, and real-time decision-making over a larger area.

References

[1] Wei, J., As'arry, A., Rezali, K., Yusoff, M., Ma, H., Zhang, K.: A review of YOLO algorithm and its applications in autonomous driving object detection. IEEE Access **13**, 93688 (2025)

[2] Wei, Y., Li, J., He, C., Yu, H., Chen, Y., Zhao, L., Wei, R.: Research progress of 3D object detection methods based on visual information and LiDAR in intelligent driving. Computer and Modernization **5**, 91 (2025)

[3] Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., Han, S.: BEVFusion: multi-task multi-sensor fusion with unified bird's-eye view representation, in Proceedings of 2023 IEEE International Conference on Robotics and Automation, London, United Kingdom, May 29-June 2 (2023), 2774-2781

[4] Khan, D., Aslam, S., Chang, K.: Vehicle-to-infrastructure multi-sensor fusion with reinforcement learning framework for enhancing autonomous vehicle perception. IEEE Access **13**, 50122 (2025)
[5] Qian, H., Wang, M., Zhu, M., Wang, H.: A review of multi-sensor fusion in autonomous driving. Sensors **25**, 6033 (2025)
[6] Liu, L.: Research on multi-sensor cooperative perception methods for intelligent driving, Ph.D. Thesis, Beijing University of Technology, School of Software Engineering (2022)
[7] 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, Glasgow, United Kingdom, August 23-28 (2020), 194-210
[8] Triva, J., Grbić, R., Vranješ, M., Teslić, N.: Weather condition classification in vehicle environment based on front-view camera images, in Proceedings of 2022 21st International Symposium INFOTEH-JAHORINA, East Sarajevo, Bosnia and Herzegovina, March 6-8 (2022), 1-4
[9] An, Z.: TransVoxelRadar: Transformer-based voxel radar perception for autonomous driving, in Proceedings of 2024 5th International Conference on Artificial Intelligence and Electromechanical Automation, Shenzhen, China, June 14-16 (2024), 1108-1112
[10] Singh, A.: Transformer-based sensor fusion for autonomous driving: a survey, in Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops, Paris, France, October 2-6 (2023), 3304-3309
[11] Wang, B., Jin, Y., Zhang, L., Zheng, L., Zhou, T.: Collaborative perception method based on multisensor fusion. Journal of Radars. **13**, 87 (2024)
[12] Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.: TransFusion: robust LiDAR-camera fusion for 3D object detection with transformers, in Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, June 18-24 (2022), 1080-1089
[13] Chen, X., Zhang, T., Wang, Y., Wang, Y., Zhao, H.: FUTR3D: A unified sensor fusion framework for 3D detection, in Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Vancouver, Canada, June 17-24 (2023), 172-181
[14] Wang, L., Zhang, X., Song, Z., Bi, J., Zhang, G., Wei, H., Tang, L., Yang, L., Li, J., Jia, C., Zhao, L.: Multi-modal 3D object detection in autonomous driving: A survey and taxonomy. IEEE Trans. Intell. Veh. **8**, 3781 (2023)
[15] Wang, P., Ma, J.: Risk2Scenario: LLM-assisted scenario generation for autonomous driving testing based on hierarchical risk analysis, in Proceedings of 2025 5th International Conference on Computer Systems, Xi'an, China, September 26-28 (2025), 55-59
[16] Wang, J., Wu, Z., Liang, Y., Tang, J., Chen, H.: Perception methods for adverse weather based on vehicle infrastructure cooperation system: A review. Sensors **24**, 374 (2024)
[17] Malik, S., Khan, M., El-Sayed, H.: Collaborative autonomous driving—A survey of solution approaches and future challenges. Sensors **21**, 3783 (2021)
[18] Chen, Y., Zhang, J., Xie, Z., Li, W., Zhang, F., Lu, J., Zhang, L.: S-NeRF++: Autonomous driving simulation via neural reconstruction and generation. IEEE Trans. Pattern Anal. Mach. Intell. **47**, 4358 (2025)