Engineering xxx (xxxx) xxx

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Agricultural Sensors—Article

Multimodal Feature Representation Mechanism for 3D Detection of Agricultural Obstacles with Few or Zero Samples

Tianhai Wang [a], Ning Wang [b], Shunda Li [a], Zhiwen Jin [b], Jianxing Xiao [a], Yanlong Miao [c], Yifan Sun [d], Han Li [a,b], Man Zhang [a,b,*]

[a] Key Laboratory of Smart Agriculture System Integration (Ministry of Education), China Agricultural University, Beijing 100083, China
[b] Key Laboratory of Agricultural Information Acquisition Technology (Ministry of Agriculture and Rural Affairs), China Agricultural University, Beijing 100083, China
[c] Key Lab of State Forestry and Grassland Administration on Forestry Equipment and Automation, School of Technology, Beijing Forestry University, Beijing 100083, China
[d] CRRC Academy, Beijing 100070, China

ARTICLE INFO

Article history:
Received 9 September 2025
Revised 25 December 2025
Accepted 5 January 2026
Available online xxxx

Keywords:
3D obstacle detection
Multimodal representation
Camera–LiDAR fusion
Autonomous navigation
Smart agriculture

ABSTRACT

Deep learning (DL) methods, particularly those that combine camera and light detection and ranging (LiDAR) data, have demonstrated remarkable accuracy in three-dimensional (3D) obstacle detection. This is crucial for achieving rigorous and reliable autonomous navigation of agricultural machinery. However, recent approaches heavily rely on large-scale labeled datasets during training, which creates challenges for their application in agriculture because of presence of scarce and distinct agricultural samples. To overcome this limitation, this paper proposes a novel 3D detection method for agricultural obstacles with few or zero samples based on a multimodal feature representation mechanism. Image and point cloud attitude adjusters are integrated to increase the accuracy, reliability, and uniformity of multimodal data. Semantic and geometry-intensity feature encoders are integrated to capture essential relationships among categories. The Bird's Eye View (BEV) fusion decoder is designed to discern intracategory similarities and intercategory distinctions. Multicategory experiments in various field scenarios reveal that the proposed method reduces the dependence on training samples by 30%–40%, and the precision rate, recall rate, F_1 score, and detection speed are 95.03%, 97.01%, 96.01%, and 16.56 frames per second (FPS), respectively. Even in completely unknown scenarios (i.e., obstacle categories that lack any corresponding training samples), the proposed method still achieves an acceptable F_1 score of 81.63%. As indicated by the results, the proposed method achieves a sophisticated trade-off among detection performance, operational efficiency, and data dependency, providing an effective safety guarantee for the autonomous navigation of agricultural machinery.

1. Introduction

Owing to the surge in population demands for food and the decrease in agricultural engagement caused by urbanization [1–3], agricultural production faces unprecedented barriers. Autonomous navigation technology, a crucial component of agricultural machinery automation, is extremely important [4], for reducing labor dependency and increasing operational efficiency [5]. However, ensuring safety during autonomous navigation remains a demanding task, particularly in preventing collisions between agricultural machinery and obstacles [6]. Robust obstacle detection is critical for guaranteeing the safe operation of autonomous systems.

Among the typical sensors in the field of obstacle detection, light detection and ranging (LiDAR) and cameras are critical components, because they provide precise point cloud and image data of the surrounding world [7]. Because single-modality data lack either color or depth information [8,9], unstructured farmland scenarios are challenging for perception systems that rely on a single sensor [10]. To address this issue, researchers have attempted collaborative solutions that exploit the complementary nature of LiDAR and camera features [11]. Camera–LiDAR-based approaches that can be jointly optimized across both modalities have rapidly evolved into a de facto criterion for robust obstacle detection [12].

* Corresponding author.
E-mail address: cauzhangman@cau.edu.cn (M. Zhang).

https://doi.org/10.1016/j.eng.2026.01.030

Please cite this article as: T. Wang, N. Wang, S. Li et al., Multimodal Feature Representation Mechanism for 3D Detection of Agricultural Obstacles with Few or Zero Samples, Engineering, https://doi.org/10.1016/j.eng.2026.01.030

1.1. Potential risks of camera–LiDAR-based solutions

Deep learning (DL) methods are powerful tools for camera–LiDAR-based obstacle detection. To process and represent multimodal data, existing solutions predominantly employ three fusion strategies: point-level fusion, feature-level fusion, and decision-level fusion [7,12,13]. The point-level fusion strategy [14–16] uses unprocessed LiDAR points to query image features, and subsequently these features are concatenated back into the point cloud as additional point-level information. The feature-level fusion strategy [17–19] begins by projecting LiDAR points into designated space or region, and subsequently associated image features are transferred back to the space or region. The decision-level fusion strategy [20–22] initially employs two single-modality detection models to obtain predicted results from various modalities, and subsequently optimizes multisource predicted results by designated fusion modules.

Although the above DL-based solutions have demonstrated remarkable efficacy, the heavy reliance on prior knowledge of obstacles and the need for extensive labeled training sets present challenges. Recent studies have highlighted that models trained on limited datasets have decreased generalizability, and the heavy reliance on prior knowledge remains challenging for reliable detection [23,24]. In unstructured farmland scenarios, prior knowledge of obstacles and large-scale labeled datasets are particularly scarce [25]. In particular, labeling a large number of samples for multimodal data is a professional and time-consuming process [26]. Therefore, introducing novel paradigms that can detect obstacles with a limited number of samples becomes more practicable and significant.

1.2. Practicable paradigms to address potential risks

Recently, few-shot and zero-shot learning paradigms have attracted considerable interest because of their ability to equip models with adaptability to new categories using minimal additional samples [27]. The objective of few-shot learning is to generalize effectively and render precise predictions for these new categories with insufficient data [28]. Zero-shot learning is a subset of few-shot learning that attempts to address new categories without any labeled samples by harnessing generalized knowledge [29]. In terms of camera-based detection, Chen et al. [30] exploited semantic interrelations among various categories in image classification tasks and executed few-shot classification via multi-tiered semantic feature enhancement. Similarly, Li et al. [31] performed few-shot image recognition by digesting a semantic-visual nexus. With respect to LiDAR-based detection, Corral-Soto et al. [32] employed the designed cycle-consistent generative adversarial network (CycleGAN) with a concurrent learning approach to increase the detection efficacy in categories with limited samples. In parallel, Li et al. [33] integrated the few-shot learning framework into point cloud detection through a sequential pair of graph neural networks. Nonetheless, the feasibility of applying few-shot and zero-shot learning paradigms within unstructured agricultural scenarios and intricate multimodal data remains to be thoroughly investigated.

1.3. Inevitable challenges in introducing practicable paradigms

Although several cases [34,35] have attempted to introduce few-shot and zero-shot learning paradigms into unstructured scenarios and multimodal data, current approaches are still not fully competent, primarily because of three inevitable challenges. First, the assurance of real-time detection speed is affected by the redundancy and complexity of multimodal data. Second, the generalized knowledge obtained from previous scenarios and categories is difficult to transfer to new farmland scenarios and categories because of sample differences caused by rugged terrain. Third, the extraction and integration of complementary features from multimodal data present considerable barriers, because the structural alignment between image and point cloud data is elusive.

1.4. Contributions of this work

On the basis of the preceding analysis, this paper proposes a novel multimodal feature representation mechanism for three-dimensional (3D) obstacle detection in unstructured agricultural scenarios. The contributions of this work are outlined as follows:

(1) To reduce unnecessary computations and guarantee real-time detection speed, a feature-level fusion strategy is used, and a voxel-based point cloud feature encoder is constructed. In contrast to the point-level fusion strategy, the feature-level fusion strategy minimizes the waste of computation in non-interest areas. Moreover, compared with point-based point cloud processing methods, the constructed voxel-based feature encoder imparts structural regularity to LiDAR data, thereby increasing computational efficiency.

(2) To minimize differences in sample structure and align various multimodal data, the BeiDou navigation satellite system (BDS) and the inertial measurement unit (IMU) are integrated and a Bird's Eye View (BEV)-based fusion decoder is constructed. The integration of the BDS and the IMU rectifies sample attitude deviations caused by rugged terrain. Furthermore, the fusion BEV decoder effectively aligns modally and structurally different data in time and space, which facilitates the subsequent discerning of intracategory similarities and intercategory distinctions based on multimodal feature representation.

(3) To mitigate the dependence of the model on labeled samples and enable 3D detection with few or zero samples, semantic and geometry-intensity feature encoders are constructed and a semantic-geometry-intensity fusion representation space is constructed. These encoders extract semantic attributes from two-dimensional (2D) image features and geometry-intensity attributes from voxel point cloud features. The extracted semantic, geometric, and intensity attributes maximize intracategory similarities and intercategory distinctions. Furthermore, the fusion representation space bridges category gaps from dimensions of semantics, geometry, and intensity.

Despite significant progress in camera–LiDAR fusion for autonomous driving, these methods presuppose extensive, well–annotated datasets—an assumption untenable in unstructured farmland, where collecting and labeling various obstacle instances is prohibitively costly and time–consuming. The proposed mechanism directly addresses this gap via feature–level fusion to focus computation on informative regions, integrating BDS and IMU data to correct field–specific attitude deviations, and unifying semantic, geometric, and intensity cues into a compact representation space. This holistic design not only minimizes training–sample requirements by 30%–40% but also maintains acceptable performance even with zero in–domain samples, providing the optimal balance among data efficiency, real–time operation, and robust generalizability needed for safe autonomous navigation in agriculture. This design provides farmers, engineers, and industry stakeholders a scalable, data–efficient, and reliable solution for safe autonomous navigation in agriculture.

2. Materials and methods

2.1. Dataset utilized for pretraining and initialization

The proposed method uses accessible public datasets to reduce the dependence on difficultly-acquired agricultural datasets.

Owing to its accessibility and sufficiency, the well-known KITTI public dataset [36] was selected to pretrain the proposed model. This dataset encompasses 7481 labeled multimodal samples in addition to 7518 unlabeled counterparts. The KITTI dataset can be downloaded directly online and remains fixed regardless of specific farmland contexts, whereas gathering and annotating 3D data for each new agricultural site requires specialized labor and is prohibitively time–consuming. By exploiting KITTI for feature initialization, the proposed method substantially reduces the dependence on costly field–specific annotations, increasing its use in various farm environments.

2.2. Obstacle detection architecture based on multimodal feature representation

In terms of a multimodal feature representation mechanism, a pioneering obstacle detection architecture is conceptualized to address the challenges stemming from the excessive reliance on large-scale labeled datasets. As shown in Fig. 1, original multiview images sequentially undergo an image attitude adjuster, an image feature encoder, and a semantic feature encoder to obtain BEV semantic features, whereas the original point clouds sequentially undergo a point cloud attitude adjuster, a point cloud feature encoder, and a geometry-intensity feature encoder to obtain BEV geometry-intensity features. The processed semantic and geometry-intensity features facilitate the transfer of generalized knowledge obtained from object categories within the KITTI dataset to obstacle categories devoid of labeled samples, because intra-category similarities and intercategory distinctions are essentially reflected in fine-grained descriptions of semantics, geometry, and intensity. The semantic and geometry-intensity features are subsequently concatenated within a unified BEV space. Ultimately, the category and position results of obstacle detection are decoded via the semantic-geometry-intensity fusion representation space and the multitask positioning head, respectively.

2.2.1. Image and point cloud attitude adjusters

Fig. 2 shows the effect of attitude deviations on obstacle perception, highlighting its detrimental effects on predicting obstacle position and orientation. Furthermore, attitude deviations affect the sequence and distribution of point clouds, increasing structural differences between KITTI and farmland samples. To maintain the accuracy, reliability, and uniformity of multimodal data, a coordinate transformation procedure is executed to eliminate attitude deviations according to the attitude information from the BDS and the IMU in the attitude adjuster. The coordinate transformation is expressed as Eq. (1):

$$R = R_Z R_X R_Y = \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\gamma & 0 & \sin\gamma \\ 0 & 1 & 0 \\ -\sin\gamma & 0 & \cos\gamma \end{bmatrix} \tag{1}$$

where R, R_Z, R_X, R_Y, ψ, θ, and γ represent the ultimate rotation matrix, rotation matrix of the yaw angle, rotation matrix of the pitch angle, rotation matrix of the roll angle, yaw angle, pitch angle, and roll angle, respectively.

2.2.2. Image and point cloud feature encoders

In terms of the image feature encoder, as shown in Fig. 1(a), the residual network (ResNet) [37] is employed as the backbone to distill preliminary multi-scale features for the trade-off between speed and accuracy. Subsequently, sequential operations including upsampling, adaptive pooling, and convolution are executed across

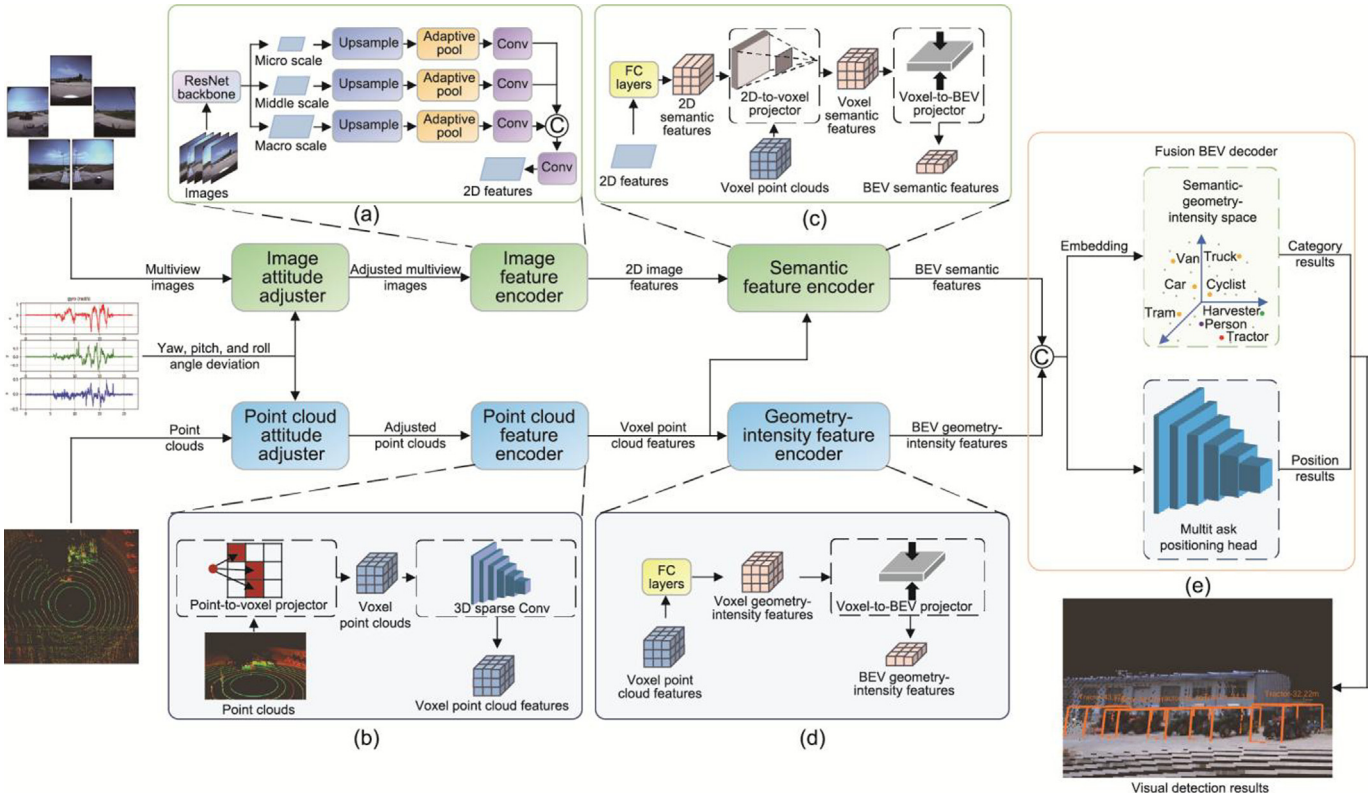

Fig. 1. Obstacle detection architecture based on multimodal feature representation. (a) Image feature encoder, (b) point cloud feature encoder, (c) semantic feature encoder, (d) geometry-intensity feature encoder, and (e) fusion BEV decoder. ResNet: residual network; FC: fully connected; ©: concatenation; Conv: convolutional layers.

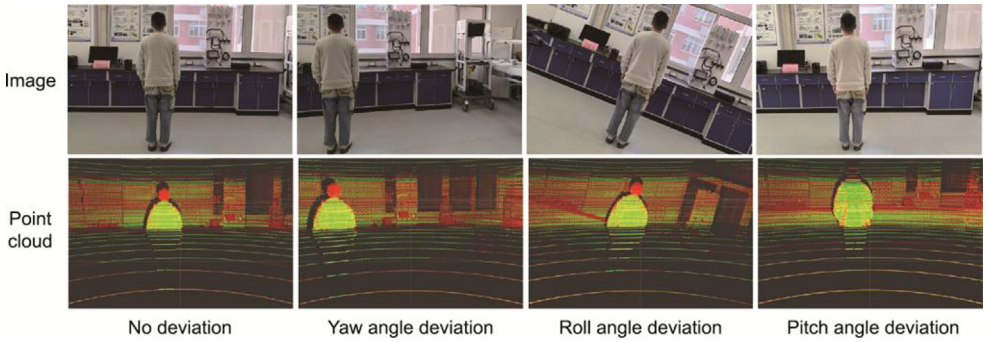

Fig. 2. Effect of attitude angle deviation on obstacle perception.

features at varying scales to capture finer details. Ultimately, a convolution operation is performed on concatenated multi-scale features to obtain refined 2D image features. The overall encoding process for images enables the capture of essential scale-invariant features.

In terms of the point cloud feature encoder, as shown in Fig. 1(b), the point-to-voxel projector segments original unordered point clouds into uniformly spaced voxel grids. Compared with unordered point clouds, voxel grids provide a more condensed representation, thereby decreasing the computational load in subsequent steps. Subsequently, 3D sparse convolutional layers are used as feature extractors for the voxel point clouds to efficiently capture the features. To address the sparsity of point clouds, sparse convolution uses interchannel and intrachannel redundancies, avoiding irrelevant computations in vacant zones.

2.2.3. Semantic and geometry-intensity feature encoders

Owing to the aptitude of fully connected (FC) layers in classification tasks [38], as shown in Figs. 1(c) and (d), FC layers are employed to encode 2D image and voxel point cloud features to semantic and geometry-intensity features. These encoded features are representations that can capture relationships among categories of interest because categories with similar semantic, geometric, and intensity attributes exhibit proximity in the vectorized space. In this work, the semantic feature encoder exclusively extracts and encodes semantic features from image features rather than from point cloud features because semantic features are reflected mainly in image data. Specifically, the semantic descriptions on which semantic features rely are typically obtained from an extensive corpus. Owing to the limitations of human sensory organs, the corpus describes mainly object attributes in terms of color. The LiDAR data are limited by the time-of-flight (ToF) perception approach, which lacks color information about the obstacles. For similar reasons, the geometry-intensity encoder solely extracts and encodes geometry and intensity features from point cloud features rather than from image features.

The projection of disparate features into a cohesive representation space without losing vital information is important. Because the transition from the 2D plane and voxel space to the BEV space is essential in the full retention of semantic features from the panoramic camera as well as geometric and intensity features from the 3D LiDAR, the BEV is used as a unified representation space for multimodal fusion. In terms of the conversion of semantic features, as shown in Fig. 1(c), the 2D-to-voxel projector voxelizes 2D semantic features that use the extrinsic calibration parameters of multi-sensors and the depth information of point clouds. The process of obtaining extrinsic calibration parameters is shown in Fig. 3, which refers to the calibration method proposed by Zhang [39]. The calibration board was placed at different distances, positions, and angles. The coordinates of the same pair of corner points in different sensor space coordinate systems can be solved to obtain the extrinsic calibration parameters. Subsequently, the voxel-to-BEV projector integrates and flattens all the features within the voxel space along the z-axis. In terms of the conversion of geometry-intensity features, as shown in Fig. 1(d), BEV geometry-intensity features are derived from their voxel counterparts via a trajectory similar to that of the semantic feature projection.

2.2.4. BEV fusion decoder

In the BEV fusion decoder, as shown in Fig. 1(e), features are concatenated in the unified BEV space to ensure the comprehensive preservation of semantic, geometric, and intensity details. Each pixel of each BEV data has multiple channels reserved to store information that may be lost during spatial transformation, such as height information. The category and position results of obstacle detection are decoded via the semantic-geometry-intensity fusion representation space and the multi-task positioning head from the BEV fusion features, respectively.

In terms of the semantic-geometry-intensity fusion representation space, the generation process is shown in Fig. 4. Semantic vectors are encoded separately from image features based on the pretrained bidirectional encoder representations from transformers (BERT) module, whereas geometry-intensity vectors are encoded together on the basis of 3D feature descriptors. Inspired by the natural language processing technique, a pretrained BERT [40] module is employed to infer semantic relationships among categories and generate semantic representation parameters. The BERT module is pretrained on an extensive corpus, such as Wikipedia and books. Furthermore, inspired by the inherent geometry and intensity details of point clouds, 3D feature descriptors, including fast point feature histograms (FPFH) and shape and intensity context combined (SICC), are used to encapsulate geometric and intensity features as high-dimensional vectors in accordance with the principles and examples outlined by Wang et al. [9]. The 3D feature descriptor is a compact and robust feature representation technique, that effectively distinguishes features by embedding relevant information into these vectors. Subsequently, the convolutional block attention module (CBAM) [41] is applied to highlight important vectors and downsize irrelevant counterparts. Vectors generated by the CBAM establish the fusion representation space, which can encapsulate semantic, geometric, and intensity relationships among categories from a mathematical perspective.

In terms of the multi-task positioning head, the center point-based strategy is used to establish the multi-task positioning head, which can facilitate the acquisition of rotational invariance and equivalence for predicting obstacles. The 3D positions of the bounding boxes are determined through the prediction of heatmaps, center offsets, and z-axis coordinates of the obstacles. The 3D size of the bounding boxes is determined through the predic-

Fig. 3. Schematic of the extrinsic parameter calibration process.

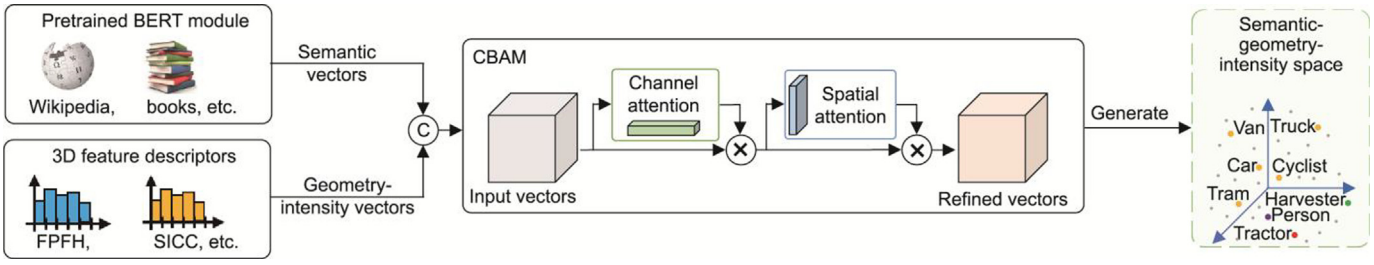

Fig. 4. Generation process of the semantic-geometry-intensity fusion representation space. CBAM: convolutional block attention module. ©: concatenation; ⊗: multiplication.

tion of the length, width, and height of the obstacles. The heading angle of the bounding boxes is determined through the prediction of the sine and cosine values of the yaw angles of the obstacles. Instead of directly predicting yaw angles, sine and cosine values of yaw angles are chosen as regression outcomes because of the remarkable continuous and periodic properties of trigonometric functions. To execute the above tasks, the multitask positioning head comprises five heads, each equipped with two convolutional layers.

2.2.5. Loss function

The overall loss comprises two components: classification loss L_{cls} and regression loss L_{reg}. For category prediction, outcomes are fine-tuned by a cross-entropy loss function because of the fundamental classification nature of this task. The classification loss for 3D bounding boxes is mathematically expressed as Eq. (2):

$$L_{cls}(P_C, G_C) = -\sum_{i=1}^{N}\sum_{j=1}^{C} G_{C_i}^{j} \log\left(P_{C_i}^{j}\right) \tag{2}$$

where P_C, G_C, N, and C represent the predicted category confidence, the ground truth (GT) of the categories, the number of total candidate results, and the number of total categories, respectively.

The cross–entropy loss in Eq. (2) is chosen to optimize the classification branch because it directly penalizes the misclassification of each candidate 3D box among the C possible obstacle categories, increasing the confidence of the model in correct semantic labels.

For position prediction, the outcomes are fine-tuned by the smooth L_1 loss function because of the fundamental regression nature of this task. The regression loss for 3D bounding boxes is mathematically expressed as Eq. (3):

$$L_{reg}(P_B, G_B) = \sum_{i=1}^{N}\sum_{k} S_{L_1}\left(P_{B_i}^{k} - G_{B_i}^{k}\right)$$

$$S_{L_1}(x) = \begin{cases} \frac{(\delta x)^2}{2}, & \text{if } |x| < \frac{1}{\delta^2}; \\ |x| - \frac{1}{2\delta^2}, & \text{otherwise.} \end{cases} \quad (3)$$

where k represents the set of the heatmap, center offset, z-axis coordinate, length, width, height, and sine and cosine values of the yaw angle; P_B and G_B represent the predicted bounding box and GT bounding box, respectively; and δ denotes a hyperparameter that controls the smoothness and is set to 1 for optimal outcomes.

The smooth L_1 loss in Eq. (3) is used for the bounding–box parameter estimation branch because it combines the robustness of L_1 loss with the stability of L_2 loss near zero error, enabling precise localization of obstacle centers, sizes, and orientations without over–penalizing occasional large deviations. The hyperparameter δ balances this trade–off. These two losses together form the overall loss, facilitating end–to–end learning of both category and position with appropriate, task–specific penalties.

3. Field experiment

3.1. Multimodal real-time perception system design

A multimodal real-time perception system was constructed on the Lovol Euro-leopard M904-D tractor, which is appropriate for unstructured farmland and rugged terrain. The architecture of the device installation is shown in Fig. 5(a). The constructed system integrates environmental perception, vehicle perception, and data fusion modules. A comprehensive overview of device types and their functions is provided in Table 1. The architecture of the synchronization data acquisition is shown in Fig. 5(b). The synchronization data acquisition system receives positioning and timing information from multiple BeiDou satellites through the BeiDou mobile station. The relevant information is decoded into real-time kinematic (RTK) position and yaw data and then transmitted to the mobile workstation at a frequency of 10 Hz. Moreover, the current time is transmitted as the time synchronization command to the mobile workstation at a frequency of 1 Hz. The multiview images and the 3D point clouds are transmitted to the mobile workstation at a frequency of 10 Hz. The pitch and roll data are transmitted to the mobile workstation at a frequency of 100 Hz. In the mobile workstation, on the basis of the robot operating system (ROS) communication framework, the current time from the BeiDou system time source is used as the medium to convert the workstation Unix time to the standard universal time coordinated (UTC) time.

3.2. Field sampling experimental design

Experiments were conducted at the Zhuozhou Experimental Station of China Agricultural University in Baoding city, Hebei Province, China, in the autumn of 2022 and the summer of 2023. Table 2 provides details of the field experimental design. The experimental scenarios comprised three typical agricultural machinery farming scenarios, namely, the cement road, the nontilled soil, and the wheat field. The experimental seasons comprised two typical busy farming seasons, including the autumn for plowing and the summer for harvesting. In terms of sampling quantity, the multimodal real-time perception system collected a total of 3029 frames of multimodal data. In terms of sampling categories, there were 2559 sample instances of the harvester category, 5407 sample instances of the tractor category, and 2108 sample instances of the person category. In terms of movement speed, the average movement speed of the tractor platform equipped with the perception system was approximately 3.6 km·h^{-1}. The movement speed of harvesters and tractors, which are typical obstacles in farmland, ranged from 3 to 8 km·h^{-1}. This speed range is consistent with the real movement speed in agricultural machinery farming environments. Considering the movement speed and the emergency braking distance of agricultural machinery, all the obstacles were set to move within the 50 m range of the perception system. If an obstacle 50 m away enters the 50 m detection range of the perception system, it is within the detection range.

Fig. 6 presents the field sampling paths. The multimodal real-time perception system started from the cement road next to the garage, and crossed the cement road to the corresponding farming scenario. In terms of the sampling in the autumn of 2022, the corresponding farming scenario was the nontilled soil scenario. In terms of the sampling in the summer of 2023, the corresponding farming scenario was the use of a wheat field. After it arrived at the corresponding farming scenario, the perception system continued to move in a straight line until it reached the end of the other side of the corresponding farming scenario. At the end of the other side, the perception system made a U-turn and then moved in a straight line back to the end of the corresponding farming scenario when entering. Throughout the sampling procedure, various obstacles (such as harvesters, tractors, and people) moved randomly within the vicinity of the perception system. In summary, the sampling path of the overall field test comprehensively reflected the real agricultural machinery work situations, which is helpful for accurately evaluating the performance of the proposed method.

Fig. 7 presents the multimodal data collected by the perception system in typical scenarios. The panoramic camera collected multiview image data through its left rear, left front, middle, right front and right rear lenses, and the 3D LiDAR collected panoramic 3D point cloud data through its 128-laser scanning section. In terms of the cement road environment, the ground was relatively flat, and multiple harvesters and tractors were parked on both sides of the road. In terms of the nontilled soil environment, the ground was uneven and soft soil, and other agricultural machines were moving randomly around the perception system. In terms of the wheat field, the terrain was uneven. A harvester was used to harvest the wheat, whereas the other tractors worked together to complete the harvest.

3.3. Implementation details and evaluation metrics

Considering that the proposed method underwent testing in real field scenarios rather than on the KITTI dataset, all labeled samples within the KITTI dataset served as valuable resources for pretraining. Limited by the high cost of annotation [42–44], in reference to related studies [45–47], key frames were extracted and labeled from the overall multimodal data stream at a frequency of approximately 1 Hz. Specifically, the number of frames in the training set, validation set, and test set collected from the field scenarios are 404, 101, and 101, respectively. The number of annotated instances in the test set is 335. The training, validation, and test sets were obtained according to the following division and augmentation techniques. Because the proposed method aims to mitigate the dependence of the model on training samples, the majority of samples from real field scenarios were not included in the training set. Specifically, total samples from real field scenarios were initially divided into training, validation, and test sets at a ratio of 1:1:1. The data augmentation technique was subsequently exclusively applied to the training set to adjust this ratio to 4:1:1. The data augmentation technique used only for the training set avoids overlap between the training set and the test set. Avoiding overlap reduces the risk of data leakage and increases the reliability of model evaluation. Training samples were gradually added to the training process to analyze the data dependency. The training process was deemed complete after the 300th epoch was reached. The Adam optimizer was employed to train the proposed model,

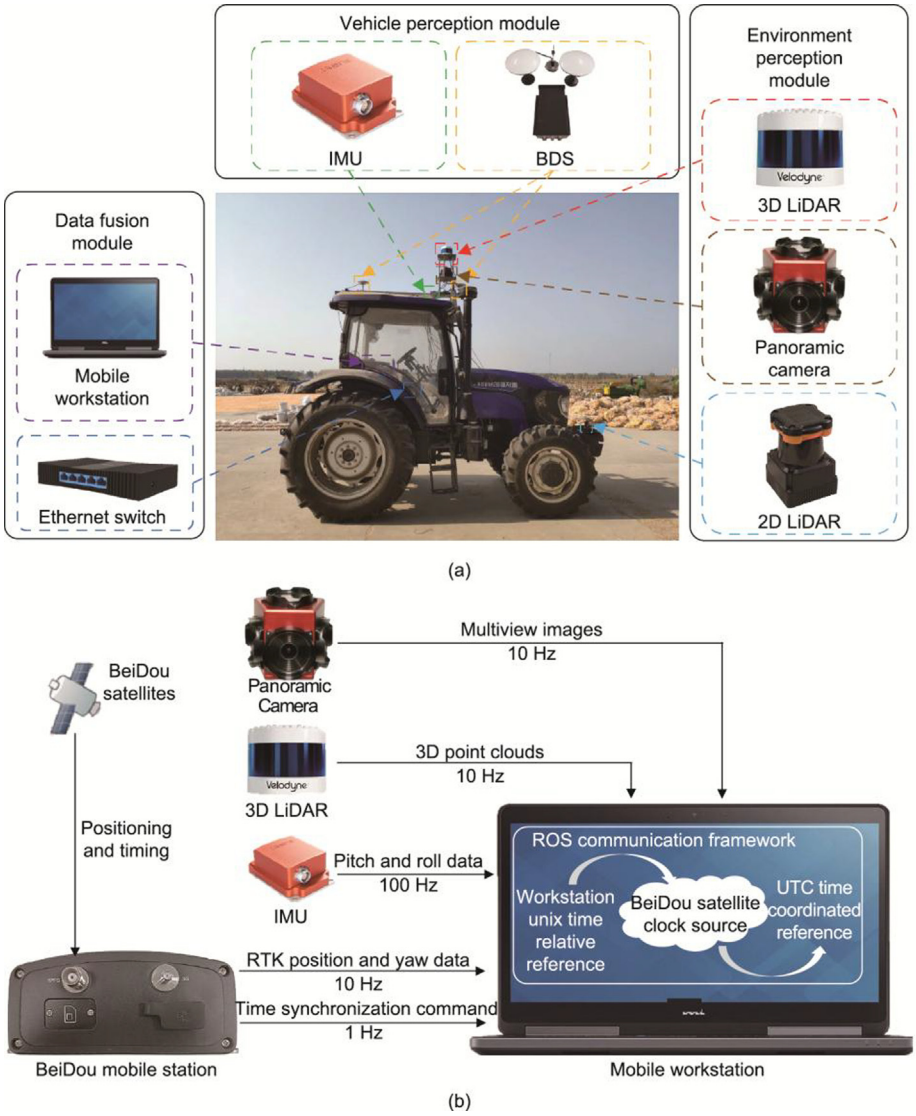

Fig. 5. Architecture of (a) device installation and (b) synchronization data acquisition.

and the cosine annealing strategy was used as the learning rate scheduler. The initial learning rate, momentum, and batch size were set to 0.001, 0.949, and 16, respectively. Finally, the model was deployed on the mobile workstation of a multimodal real-time perception system and tested in real field scenarios.

Samples can be classified into four types according to the combination of labeled results and predicted results for binary classification: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). According to related studies [48–50], the TP of the harvester and tractor categories is defined as the predicted category being correct and the value of the 3D intersection over union (IoU) being greater than or equal to 0.7. Moreover, the TP of the person category is defined as the predicted category being correct and the value of the 3D IoU being greater than or equal to 0.5. 3D IoU [51] measures the overlap between two 3D volumes, such as predicted and GT bounding boxes, and is defined as the ratio of the volume of their intersection to the volume of their union. It is mathematically delineated as Eq. (4):

$$\mathrm{IoU}_{3D}(B_P, B_G) = \frac{\mathrm{Volume}\ (B_P \cap B_G)}{\mathrm{Volume}\ (B_P \cup B_G)} \tag{4}$$

where B_P and B_G represent the predicted 3D bounding box and the GT 3D bounding box, respectively.

Recall and precision are applied to evaluate the different detection performances. Recall is the proportion of TP results among the results that are actually positive, whereas precision is the proportion of TP results among the results that are predicted to be positive. Intuitively, the recall reveals the ability of a model to correctly predict all positive samples, whereas the precision reveals the ability of a model not to predict negative samples as positive samples. Recall and Precision are mathematically expressed as Eqs. (5) and (6), respectively:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

where TP, FP, TN, and FN represent the number of TPs, FPs, TNs, and FNs, respectively.

Recall and Precision usually exhibit a trade-off relationship. To comprehensively assess the performance, the F_1 score serves as

Table 1
Comprehensive overview of device types and their functions.

Device	Models	Functions
3D LiDAR	VLS-128 (Velodyne, USA)	• Gathers 3D point clouds within 360°
Panoramic camera	Ladybug 5 (PointGrey, Canada)	• Gathers multi-view images within 360°
2D LiDAR	UTM-30LX-EW (HOKUYO, Japan)	• Covers 3D LiDAR blind spots for emergency stop • Not used for this work
BeiDou base station	I60 (CHCNAV, China)	• Receives position data from satellites • Computes deviations between measured and true values • Transmits position deviations via radio communication
BeiDou mobile station	P3-DT (CHCNAV, China)	• Receives position data from satellites • Receives position deviations via radio communication • Computes the RTK position and heading based on dual antennas
IMU	MTi-300 (Xsens, Netherlands)	• Monitors real-time attitude deviations of the system • Generates rotation matrices to correct attitude deviations
Mobile workstation	Y9000P 2021 (Lenovo, China)	• Establishes communication with all sensors to obtain data • Implements the obstacle detection model
Ethernet switch	TL-SG1005M (TP-Link, China)	• Extends the physical interface to enable simultaneous communication with multiple devices

Table 2
Details of the field experimental design.

Name	Value
Scenarios	Cement road, nontilled soil, wheat field
Seasons	Autumn, summer
Overall multimodal frames	3029
Categories of obstacles	Harvester (2559 instances) Tractor (5407 instances) Person (2108 instances)
Constant speed of the real-time perception system	Approximately 3.6 km·h^{-1}
Constant speed of the harvesters and tractors	3–8 km·h^{-1}
Constant speed of the people	Approximately 3.6 km·h^{-1}
Maximum detection range	50 m

the harmonic mean of Recall and Precision and is mathematically expressed as Eq. (7):

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

The frame rate is the reciprocal of the average processing time per frame, which is used to assess the detection speed and is mathematically expressed as Eq. (8):

$$\text{FrameRate} = \frac{N}{\sum_{i=1}^{N} t_i} \tag{8}$$

where N and t_i represent the number of total frames and the processing time of each frame, respectively.

In summary, the 3D IoU is a metric for evaluating the regression task. The recall, precision, and F_1 score are comprehensive metrics that consider both classification and regression tasks. The frame rate is a metric for evaluating detection speed.

4. Results and discussion

4.1. Performance comparison and data dependency analysis

The proposed method is compared with four state-of-the-art methods for 3D obstacle detection, namely, BEVFusion [7], multimodal voxel net (MVXNet) [14], point-voxel region-based convolutional neural network plus plus (PV-RCNN++) [52], and flexible monocular 3D object detection (MonoFlex) [53]. The proposed method, BEVFusion, and MVXNet are camera-LiDAR-based solutions, whereas the PV-RCNN++ and MonoFlex are LiDAR-based and camera-based solutions, respectively. As shown in Fig. 8, at 100% usage of the training set, the overall F_1 score achieved by the proposed method is 3.78% higher than that of the baseline method (i.e., BEVFusion). The demand and dependence of different methods on the training set are quantitatively characterized by the usage rate of the training set. For the purpose of detection with few samples, at the same overall F_1 score, the proposed method decreases the usage of the training set by 30%–40% compared with the baseline method, which shows that the proposed method effectively mitigates the dependence of the model on extensive labeled training samples. For the purpose of detection with zero samples, without any training samples (i.e., the usage rate of the training set is 0), the precision rate, recall rate, and F_1 score of the proposed method are guaranteed to be approximately 80%, which proves that the proposed method can still provide effective safety guarantees for autonomous navigation of agricultural machinery even in extremely unknown scenarios.

As shown in Fig. 8(a), an intriguing pattern is observed in the precision rate curves of the alternative methods: they initially increase, but then decrease, and finally increase again. This phenomenon contradicts intuitive expectations and may stem from the pretraining process on the KITTI dataset, which tends to bias model decisions toward conservatism. While a higher precision rate accompanied by a lower recall rate may seem to indicate acceptable performance, it poses a significant risk. For example, in the most conservative case (i.e., the model does not generate any bounding boxes for any obstacles), the precision rate may approach 100%, but the recall rate decreases to nearly 0. Therefore, a comprehensive evaluation, incorporating different perspectives via multiple evaluation metrics, is imperative to accurately reflect performance. Because the designed semantic-geometry-intensity fusion representation space improves the ability to identify intra-category similarities and intercategory distinctions when samples are extremely scarce, the proposed method achieves acceptable performance at different usage rates of the training set.

4.2. Detection results of the proposed method across various categories

Fig. 9 presents the detection results of the proposed method across various categories. In general, the proposed method exhibits

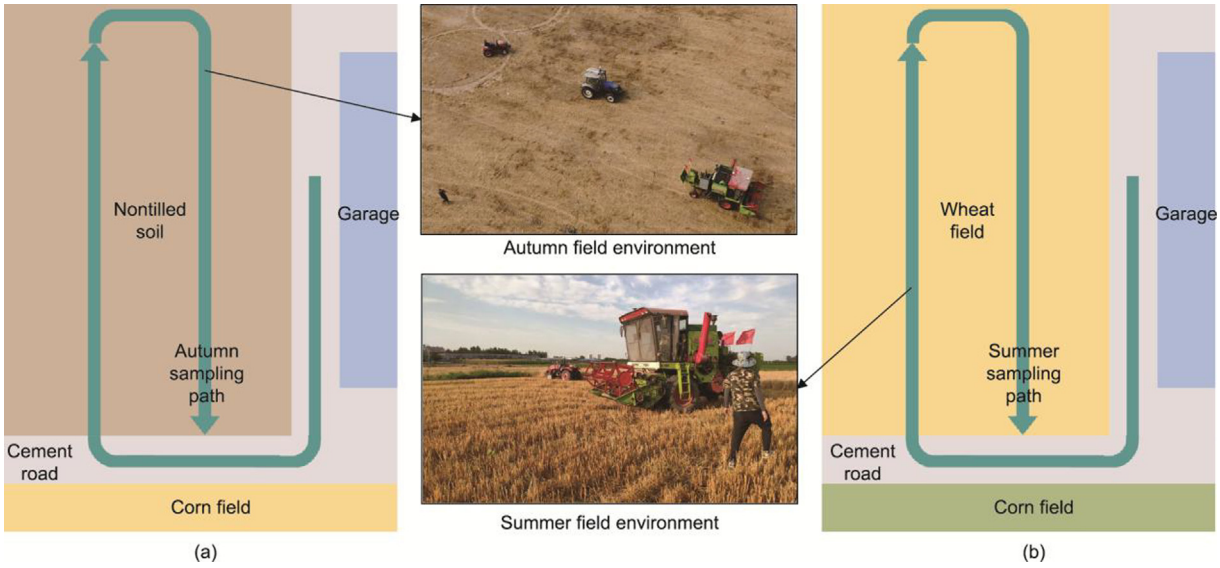

Fig. 6. Illustration of field sampling paths: (a) autumn 2022 and (b) summer 2023.

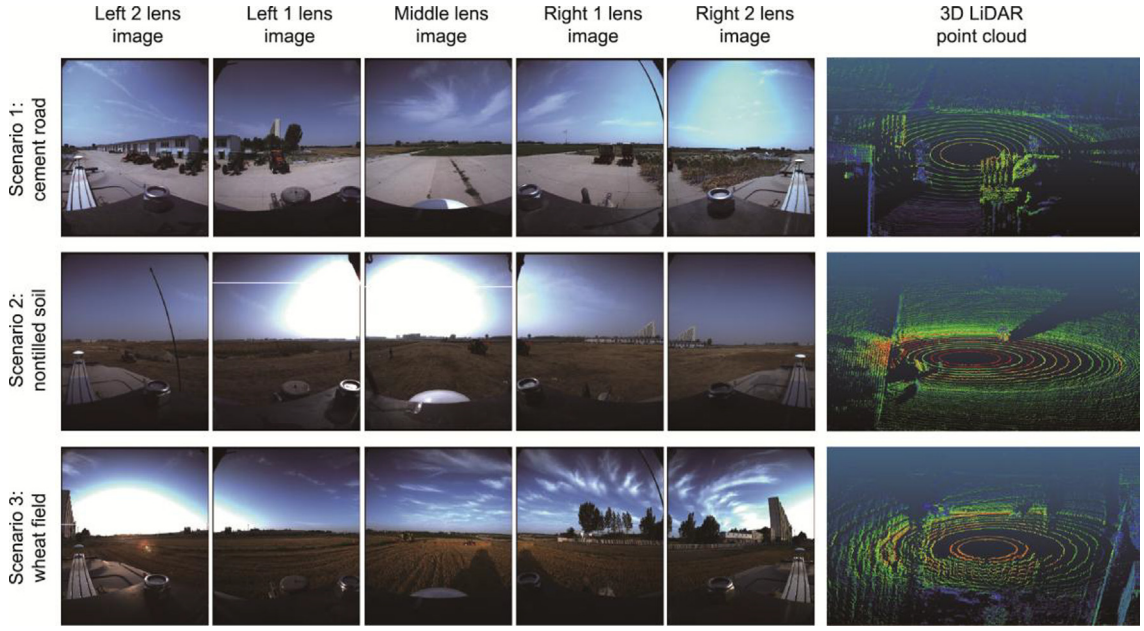

Fig. 7. Multimodal data collected in typical scenarios.

commendable performance for all categories, attaining an overall F_1 score of 96.01%. Notably, the precision rate, recall rate, and F_1 score of the proposed method for both the harvester and tractor categories exceed 95%.

As shown in Fig. 9, the precision rate, recall rate, and F_1 score of the proposed method for tractors and harvesters exceed 95%, whereas those for people are nearly 95%. This finding is attributed to the different number of pixels and points of instances in various categories. Specifically, the 3D size of the person category is smaller, resulting in fewer pixels and points within multimodal data. Considering the multimodal fusion structure in the form of colorized point clouds as the statistical subject, the statistical analysis reveals that each instance of harvesters comprises an average of 5849 points, each instance of tractors comprises an average of 785 points, and each instance of people comprises only 181 points on average. Furthermore, compared with large objects, small objects are more susceptible to the negative effect of registration errors under the same registration accuracy. In terms of the 3D IoU metric, the 3D IoU value of the proposed method for all categories is greater than 75%, which is greater than the 3D IoU value of 70% (i.e., 0.7) commonly set for obstacle detection in the field of autonomous driving [54–56]. The 3D IoU value is partially affected by the accuracy of the GT bounding box. Owing to the lack of geometric information about the occluded components of obstacles, the geometric size of the GT bounding boxes must be manually set through experience when dealing with partially occluded obstacles.

4.3. Detection efficiency of the proposed method

Efficient real-time obstacle detection is crucial because faster detection requires more time and space for avoidance maneu-

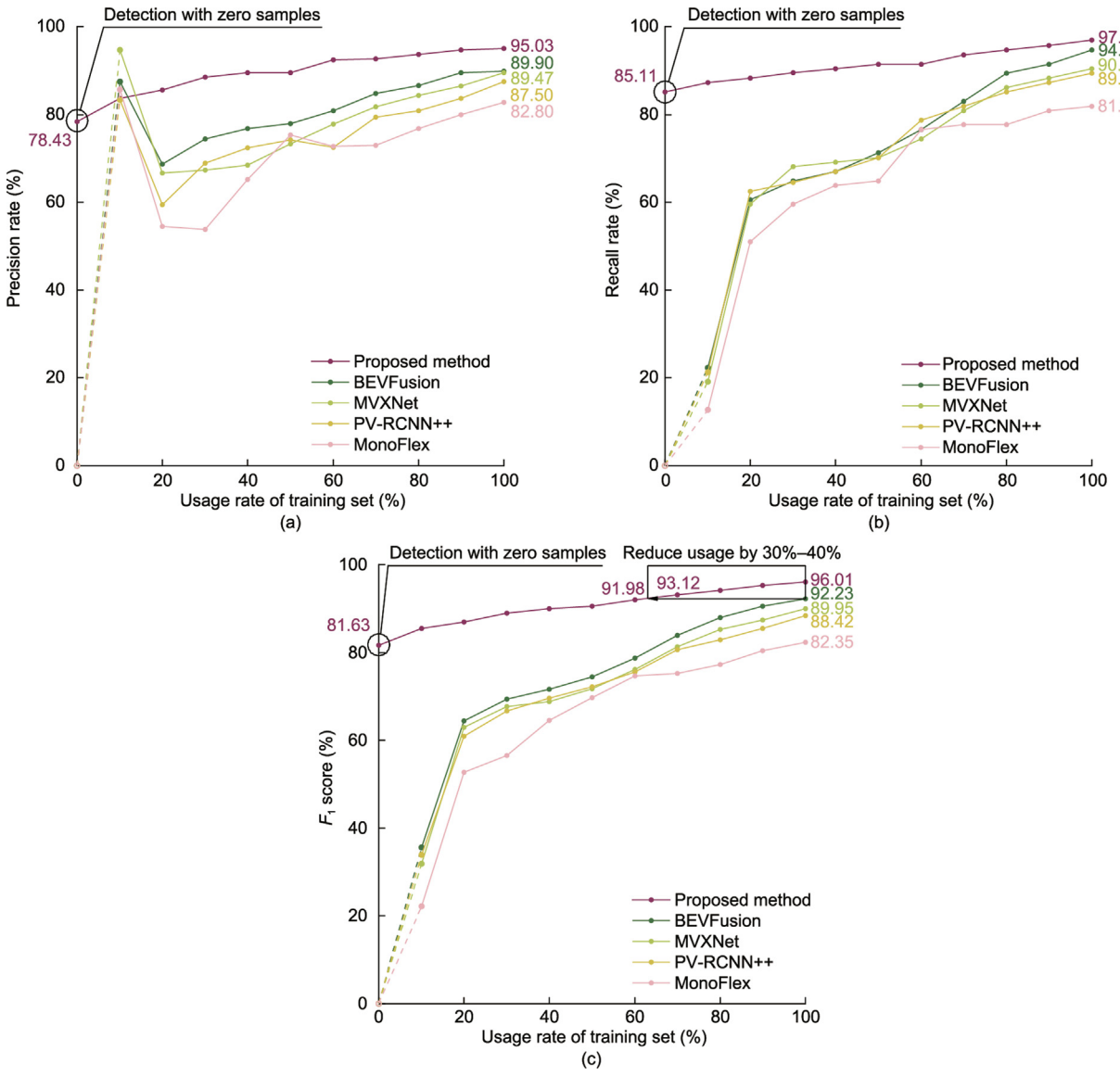

Fig. 8. Comparison of performance under different usage rates of the training set: (a) precision rate, (b) recall rate, and (c) F_1 score.

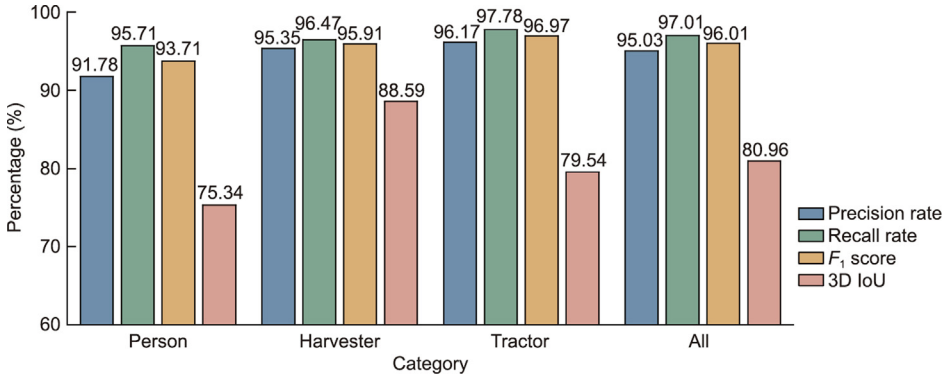

Fig. 9. Comparison of the detection results of the proposed method across various categories.

vers. Because of the inherent spatial and energy limitations of agricultural machinery, the detection efficiency obtained on mobile workstations has greater representational value than that on their desktop counterparts. As shown in Table 3, the proposed method achieves a frame rate of 16.56 frames per second (FPS), exceeding the sampling rates of both the 3D LiDAR and the panoramic camera employed in this work. Because the processing speed outpaces the sampling speeds of the sen-

Table 3
Real-time performance comparison according to the processing and sampling speeds.

Method/device	Time consumption per frame (millisecond)	Frame rate (FPS)
3D LiDAR	100	10
Panoramic camera	100	10
Proposed method	60.39	16.56

sors, the proposed method is deemed to satisfy the real-time demand of obstacle detection.

4.4. Visual detection results for different obstacles in various scenarios

To emphasize the superior performance of the proposed method, the visual detection results of the proposed method in three typical scenarios are shown in Fig. 10. The proposed method generates well-fitted 3D bounding boxes and fine-grained categories for different obstacles encountered in various scenarios.

In most visual detection results, to reduce computing power consumption, only the point clouds within the maximum detection range of 100 m of the real-time data acquisition system are assigned the color of the corresponding pixel position. The point clouds beyond the range of 100 m are directly filtered out, so the middle and upper components of each visual detection result are black.

In some visual detection results, the colorized point clouds in the area near the vehicle body are sparse, and the colorized point clouds in the area far from the vehicle body are dense. This is because the 128-laser scanning section of the 3D LiDAR is not linearly distributed in the vertical field of view. Specifically, the overall layout is dense in the middle and sparse on the upper and lower sides. Reflected in visual detection results of the colorized point clouds, the points in the middle are dense, whereas the points in the lower part are sparse. In particular, the upper part is black because it exceeds the detection range.

A few visual detection results (e.g., Fig. 10(f)) reveal some white and light blue point clusters. This is because the installation positions of the 3D LiDAR and the panoramic camera are different, and parallax between the two sensors is observed. When a component of the 3D LiDAR is blocked but the corresponding area of the panoramic camera is not blocked, the 3D point cloud of the corresponding area may be mistakenly matched to the color behind the blocker, such as the sky. This phenomenon does not have a significant negative effect on the detection results.

5. Conclusions

In this paper, a novel 3D detection method for obstacles with few or zero corresponding samples is proposed on the basis of a

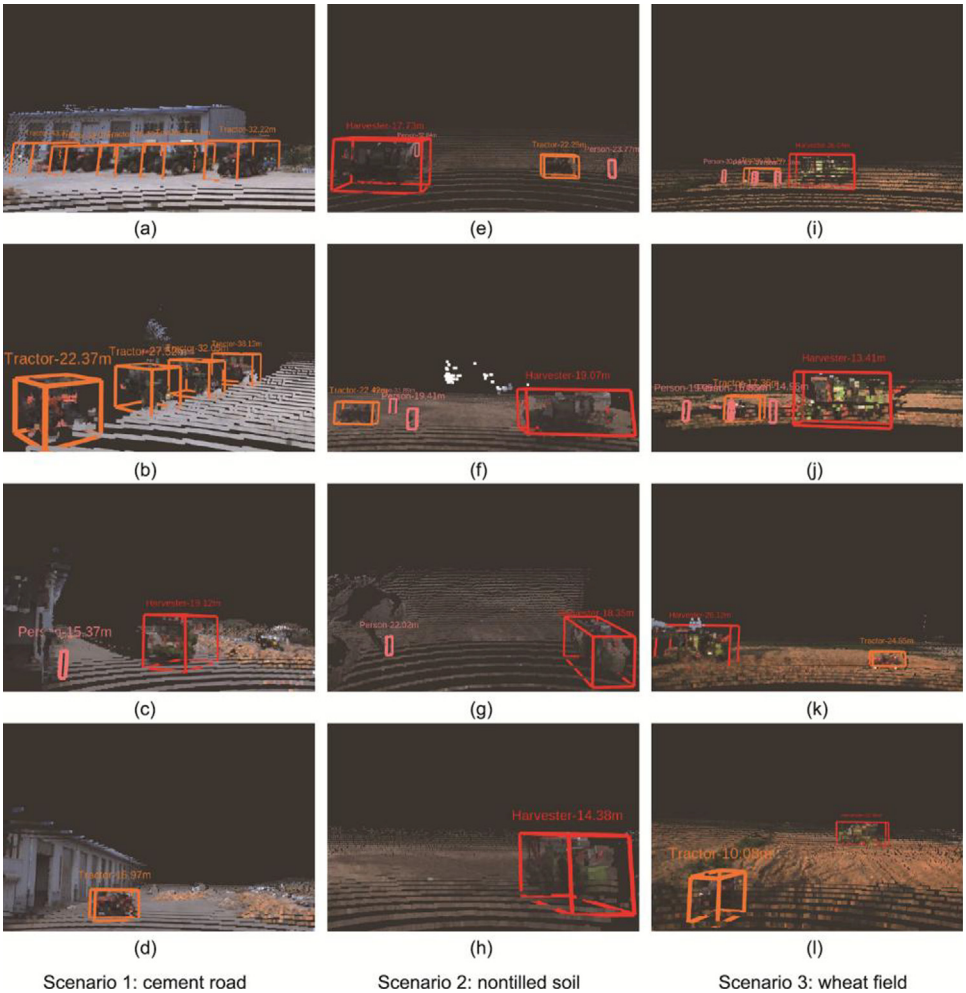

Fig. 10. Visual detection results of the proposed method in typical scenarios. (a–d) Cement road, (e–h) nontilled soil, (i–l) wheat field.

multimodal feature representation mechanism. Eliminating attitude deviations according to attitude information from the BDS and the IMU increases the accuracy, reliability, and uniformity of multimodal data. By encoding semantic, geometric, and intensity features from original multimodal data, essential relationships among categories are captured. By identifying intracategory similarities and intercategory distinctions via a semantic-geometry-intensity fusion representation space, the proposed method effectively bridges category gaps with a limited number of labeled samples. Real field experiments reveal that the proposed method reduces the dependence on training samples by 30%–40%, and the precision rate, recall rate, F_1 score, and detection speed are 95.03%, 97.01%, 96.01%, and 16.56 FPS, respectively. Even in completely unknown scenarios (i.e., obstacle categories that lack any corresponding training samples), the proposed method still maintains an acceptable F_1 score of 81.63%.

While the results prove that the proposed method achieves an encouraging trade-off among detection performance, operational efficiency, and data dependency, several limitations still exist. In this work, the extrinsic parameters of multi-sensors are precalibrated and fixed and may not be robust enough to obstacles with variable scales. Future work will explore scale-adaptive multimodal registration strategies to improve the generalizability for multiscale obstacles.

CRediT authorship contribution statement

Tianhai Wang: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Ning Wang:** Validation, Investigation, Formal analysis, Data curation. **Shunda Li:** Formal analysis, Data curation. **Zhiwen Jin:** Validation, Investigation, Data curation. **Jianxing Xiao:** Validation, Investigation, Data curation. **Yanlong Miao:** Software, Investigation. **Yifan Sun:** Investigation. **Han Li:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Man Zhang:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2022YFD2001600-2022YFD2001601).

References

[1] Zhang L, Huang L, Li T, Wang T, Yang X, Yang Q. The skyscraper crop factory: a potential crop-production system to meet rising urban food demand. Engineering 2023;31:70–5.
[2] Zhang S, Chen Z, Cao C, Cui Y, Gao Y. Photothermal-management agricultural films toward industrial planting: opportunities and challenges. Engineering 2023;35:191–200.
[3] Wang T, Liu Y, Wang M, Fan Q, Tian H, Qiao X, et al. Applications of UAS in crop biomass monitoring: a review. Front Plant Sci 2021;12:616689.
[4] Wei W, Xiao M, Wang H, Zhu Y, Xie C, Geng G. Research progress of multiple agricultural machines for cooperative operations: a review. Comput Electron Agric 2024;227:109628.
[5] Yu Y, Liu Y, Wang J, Noguchi N, He Y. Obstacle avoidance method based on double DQN for agricultural robots. Comput Electron Agric 2023;204:107546.
[6] Wang J, Zheng H, Yu Y, He Y, Liu Y. Robust multiple obstacle tracking method based on depth aware OCSORT for agricultural robots. Comput Electron Agric 2024;217:108580.
[7] Liu Z, Tang H, Amini A, Yang X, Mao H, Rus DL, et al. BEVFusion: multi-task multi-sensor fusion with unified bird's-eye view representation. In: Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA); 2023 May 29–Jun 2; London, UK. New York City: IEEE; 2023. p. 2774–81.
[8] Wang T, Chen B, Wang N, Ji Y, Li H, Zhang M. Zero-shot obstacle detection using panoramic vision in farmland. J Field Robot 2024;41(7):2169–83.
[9] Wang T, Wang N, Xiao J, Miao Y, Sun Y, Li H, et al. One-shot domain adaptive real-time 3D obstacle detection in farmland based on semantic-geometry-intensity fusion strategy. Comput Electron Agric 2023;214:108264.
[10] Li T, Xie F, Feng Q, Qiu Q. Multi-vision-based localization and pose estimation of occluded apple fruits for harvesting robots. In: Proceedings of the 2022 37th Youth Academic Annual Conference of Chinese Association of Automation (YAC); 2022 Nov 19–20; Beijing, China. New York City: IEEE; 2022. p. 767–72.
[11] Wei W, Xiao M, Duan W, Wang H, Zhu Y, Zhai C, et al. Research progress on autonomous operation technology for agricultural equipment in large fields. Agriculture 2024;14(9):1473.
[12] Liang T, Xie H, Yu K, Xia Z, Lin Z, Wang Y, et al. BEVFusion: a simple and robust LiDAR-camera fusion framework. In: Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022); 2022 Nov 28–Dec 9; New Orleans, LA, USA. Red Hook: Curran Associates, Inc.; 2022. p. 10421–34.
[13] Wang X, Li K, Chehri A. Multi-sensor fusion technology for 3D object detection in autonomous driving: a review. IEEE Trans Intell Transp Syst 2024;25(2):1148–65.
[14] Sindagi VA, Zhou Y, Tuzel O. MVX-Net: multimodal VoxelNet for 3D object detection. In: Proceedings of the 2019 IEEE International Conference on Robotics and Automation (ICRA); 2019 May 20–24; Montreal, QC, Canada. New York City: IEEE; 2019. p. 7276–82.
[15] Wang C, Ma C, Zhu M, Yang X. PointAugmenting: cross-modal augmentation for 3D object detection. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 19–25; Virtual Conference. Los Alamitos: IEEE Computer Society; 2021. p. 11789–98.
[16] Xie B, Yang Z, Yang L, Wei A, Weng X, Li B. AMMF: attention-based multi-phase multi-task fusion for small contour object 3D detection. IEEE Trans Intell Transp Syst 2022;24:1692–701.
[17] Zhao K, Ma L, Meng Y, Liu L, Wang J, Junior JM. 3D vehicle detection using multi-level fusion from point clouds and images. IEEE Trans Intell Transp Syst 2022;23(9):15146–54.
[18] Chen Y, Li Y, Zhang X, Sun J, Jia J. Focal sparse convolutional networks for 3D object detection. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. Los Alamitos: IEEE Computer Society; 2022. p. 5418–27.
[19] Bai X, Hu Z, Zhu X, Huang Q, Chen Y, Fu H, et al. TransFusion: robust LiDAR-camera fusion for 3D object detection with transformers. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. Los Alamitos: IEEE Computer Society; 2022. p. 1080–9.
[20] Shi P, Qi H, Liu Z, Yang A. 3D vehicle detection algorithm based on multimodal decision-level fusion. Comput Model Eng Sci 2023;135(3):2007–23.
[21] Pang S, Morris D, Radha H. CLOCs: camera-LiDAR object candidates fusion for 3D object detection. In: Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2020 Oct 25–29; Las Vegas, NV, USA. New York City: IEEE; 2020. p. 10386–93.
[22] Ku J, Mozifian M, Lee J, Harakeh A, Waslander SL. Joint 3D Proposal Generation and Object Detection from View Aggregation. In: Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2018 Oct 1–5; Madrid, Spain. New York City: IEEE; 2018. p. 1–8.
[23] Zhang Q, Barri K, Babanajad SK, Alavi AH. Real-time detection of cracks on concrete bridge decks using deep learning in the frequency domain. Engineering 2021;7(12):1786–96.
[24] Wang H, Ma Z, Ren Y, Du S, Lu H, Shang Y, et al. Interactive image segmentation based field boundary perception method and software for autonomous agricultural machinery path planning. Comput Electron Agric 2024;217:108568.
[25] Wang T, Chen B, Zhang Z, Li H, Zhang M. Applications of machine vision in agricultural robot navigation: a review. Comput Electron Agric 2022;198:107085.
[26] Tian H, Wang T, Liu Y, Qiao X, Li Y. Computer vision technology in agricultural automation—a review. Inf Process Agric 2020;7(1):1–19.
[27] Wang H, Li J, Wu H, Hovy E, Sun Y. Pre-trained language models and their applications. Engineering 2023;25:51–65.
[28] Wang X, Wang X, Jiang B, Luo B. Few-shot learning meets transformer: unified query-support transformers for few-shot classification. IEEE Trans Circ Syst Video Tech 2023;33(12):7789–802.
[29] Gupta A, Narayan S, Khan S, Khan FS, Shao L, van de Weijer J. Generative multi-label zero-shot learning. IEEE Trans Pattern Anal Mach Intell 2023;45(12):14611–24.
[30] Chen Z, Fu Y, Zhang Y, Jiang YG, Xue X, Sigal L. Multi-level semantic feature augmentation for one-shot learning. IEEE Trans Image Process 2019;28(9):4594–605.
[31] Li Z, Tang H, Peng Z, Qi GJ, Tang J. Knowledge-guided semantic transfer network for few-shot image recognition. IEEE Trans Neural Netw Learn Syst 2025;36(11):19474–88.
[32] Corral-Soto ER, Nabatchian A, Gerdzhev M, Bingbing L. LiDAR few-shot domain adaptation via integrated CycleGAN and 3D object detector with joint learning delay. In: Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA); 2021 May 30–Jun 5; Xi'an, China. New York City: IEEE; 2021. p. 13099–105.

[33] Li Y, Chen C, Yan W, Cheng Z, Tan HL, Zhang W. Cascade graph neural networks for few-shot learning on point clouds. IEEE Trans Intell Transp Syst 2023;24(8):8788.
[34] Hong B, Zhou Y, Qin H, Wei Z, Liu H, Yang Y. Few-shot object detection using multimodal sensor systems of unmanned surface vehicles. Sensors 2022;22(4):1511.
[35] Zhu K, Chen W, Hou Z, Wang Q, Chen H. Modified fusing-and-filling generative adversarial network–based few-shot image generation for GMAW defect detection using multi-sensor monitoring system. Int J Adv Manuf Technol 2023;128(5–6):2753–62.
[36] Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: the KITTI dataset. Int J Rob Res 2013;32(11):1231–7.
[37] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. Los Alamitos: IEEE Computer Society; 2016. p. 770–8.
[38] Wu Y, Chen Y, Yuan L, Liu Z, Wang L, Li H, et al. Rethinking classification and localization for object detection. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. Los Alamitos: IEEE Computer Society; 2020. p. 10183–92.
[39] Zhang Z. A flexible new technique for camera calibration. IEEE Trans Pattern Anal Mach Intell 2000;22(11):1330–4.
[40] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019); 2019 Jun 2–7; Minneapolis, MN, USA. Stroudsburg: Association for Computational Linguistics; 2019. p. 4171–86.
[41] Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Proceedings of the Computer Vision—ECCV 2018: 15th European Conference; 2018 Sep 8–14; Munich, Germany. Cham: Springer International Publishing; 2018. p. 3–19.
[42] Wang H, Shan Y, Chen L, Liu M, Wang L, Meng Z. Multi-scale feature learning for 3D semantic mapping of agricultural fields using UAV point clouds. Int J Appl Earth Obs Geoinf 2025;141:104626.
[43] Wu A, He P, Li X, Chen K, Ranka S, Rangarajan A. An efficient semi-automated scheme for infrastructure LiDAR annotation. IEEE Trans Intell Transp Syst 2024;25(7):8237–47.
[44] Gao B, Pan Y, Li C, Geng S, Zhao H. Are we hungry for 3D LiDAR data for semantic segmentation? A survey of datasets and methods. IEEE Trans Intell Transp Syst 2022;23(7):6063–81.
[45] Guan X, Wan H, Han W, Jiang R, Ou Y, Chen Y, et al. MDS-PointPillars: a lightweight obstacle identification method in farmland based on three-dimensional LiDAR for autonomous navigation. Comput Electron Agric 2025;237:110688.
[46] Wu T, Guo H, Zhou W, Gao G, Wang X, Yang C. Navigation path extraction for farmland headlands via red-green-blue and depth multimodal fusion based on an improved DeepLabv3+ model. Eng Appl Artif Intel 2025;151:110681.
[47] Jiang W, Chen W, Song C, Yan Y, Zhang Y, Wang S. Obstacle detection and tracking for intelligent agricultural machinery. Comput Electr Eng 2023;108:108670.
[48] Chi F, Wang Y, Nasiopoulos P, Leung VCM. Parameter-efficient federated cooperative learning for 3-D object detection in autonomous driving. IEEE Internet Things J 2025;12(12):20314–25.
[49] Wang Z, Li YL, Chen X, Zhao H, Wang S. Uni3DETR: unified 3D detection transformer. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023 (NeurIPS 2023); 2023 Dec 10–16; New Orleans, LA, USA. Red Hook: Curran Associates, Inc.; 2023. p. 39876–96.
[50] Jin R, Jia Z, Chu Z. Efficient aerial image object detection with imaging condition decomposition. In: Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP 2023); 2023 Oct 8–11; Kuala Lumpur, Malaysia. New York City: IEEE; 2023. p. 620–4.
[51] Ravi N, Reizenstein J, Novotny D, Gordon T, Lo WY, Johnson J, et al. Accelerating 3D Deep Learning with PyTorch3D. arXiv:2007.08501.
[52] Shi S, Jiang L, Deng J, Wang Z, Guo C, Shi J, et al. PV-RCNN++: point-voxel feature set abstraction with local vector representation for 3D object detection. Int J Comput Vis 2023;131(2):531–51.
[53] Zhang Y, Lu J, Zhou J. Objects are different: flexible monocular 3D object detection. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021); 2021 Jun 20–25; Nashville, TN, USA. Los Alamitos: IEEE Computer Society; 2021. p. 3288–97.
[54] Lu B, Sun Y, Yang Z, Song R, Jiang H, Liu Y. HRNet: 3D object detection network for point cloud with hierarchical refinement. Pattern Recogn 2024;149:110254.
[55] Chen H, Yan H, Yang X, Su H, Zhao S, Qian F. Efficient adversarial attack strategy against 3D object detection in autonomous driving systems. IEEE Trans Intell Transp Syst 2024;25(11):16118–32.
[56] Sheng H, Cai S, Zhao N, Deng B, Huang J, Hua XS, et al. Rethinking IoU-based optimization for single-stage 3D object detection. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. Proceedings of the Computer Vision—ECCV 2022: 17th European Conference; 2022 Oct 23–27; Tel Aviv, Israel.