Multimodal Perception and Real-Time Fusion of Biomimetic Robots

Yifei Xu

{marineee831@gmail.com}

Faculty of Intelligent Technology, Shanghai Institute of Technology, Shanghai, 201400, China

Abstract. Perception in intelligent robots is shifting away from single-modality sensing toward the coordinated use of multiple information channels. Focusing on intelligent bionic robots, this paper reviews how bionics and embodied intelligence have shaped the joint design of sensing hardware and perceptual algorithms. Then further investigates three representative paradigms for multimodal information fusion, namely point-cloud fusion, heterogeneous fusion, and bird's-eye-view (BEV) fusion. In response to current limitations, especially insufficient device-level integration and weak cross-scene generalization, this paper provided a future development route centered on integrated hardware that combines communication, sensing, and transmission, together with optoelectronic chips and generative artificial intelligence models, in order to improve overall system performance and adaptive capability. The study is intended to provide a systematic reference for the advancement of multimodal perception in intelligent bionic robots and to support further progress toward stronger environmental understanding and more capable autonomous decision-making.

Keywords: Intelligent bionic robots; Multimodal perception; Perception technology

1 Introduction

The development of intelligent robots is fundamentally tied to the evolution of their perceptual capabilities. Early robotic systems relied on a single sensing modality, such as vision or force sensing, to interpret their surroundings. Although such systems were sufficiently adaptable in structured environments, they often struggled to function effectively in complex, dynamic, and unstructured settings. As robotic tasks have become increasingly sophisticated and bionic principles have been introduced into robot design, perceptual strategies have gradually shifted from single-modality sensing to the fusion of multiple sources of information. Bionic robots, by drawing inspiration from biological morphology and behavioral mechanisms, exhibit clear advantages in both environmental adaptability and interactive performance. At the same time, the rise of embodied intelligence has accelerated the application of multimodal large-model approaches to core robotic functions such as perception and decision-making, while also highlighting the growing importance of tightly coupled perception-action systems. Against

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365332

this backdrop, perceptual hardware is likewise moving toward greater integration and intelligence. This achievement itself depended on the multimodal acquisition and fusion of biological motion data, and it represented an early attempt to transform biological perception-motion mechanisms into embodied perception-motion capabilities in robots [1]. Taken together, the field has progressed from single-sensor information processing to multisource information acquisition under bionic morphological frameworks, and further toward the closed-loop integration of perception, actuation, and decision-making based on intelligent materials and algorithms. In this context, embodied multimodal perception has become a core enabling technology for advancing intelligent bionic robots to a higher level of capability.

Multimodal perception lies at the heart of how intelligent systems interact with their environments. In recent years, it has advanced substantially in both theoretical research and practical application, showing a clear trend toward interdisciplinary expansion and increasing depth, with its scope extending from industrial automation to operations in extreme environments. Kong et. al reported that, in the operation and maintenance of intelligent power stations for new energy systems, multimodal fusion perception has been used to build an all-dimensional sensing framework for inspection robot dogs through the integration of visual, infrared, acoustic, and millimeter-wave radar data [2]. By overcoming the informational limitations of individual sensors in complex environments, this framework provides a representative technical pathway for enabling fully autonomous operation of mobile robots in unstructured scenarios. In the area of precise robotic manipulation, tactile perception is a key enabling technology for dexterous grasping and safe human-robot interaction. A systematic review of tactile sensor principles, simulation platforms, and representative application tasks can therefore offer important theoretical guidance for the continued improvement and integration of multimodal perception systems from the perspective of touch. Moreover, multimodal perception is increasingly being extended to the monitoring of living states. Lin argued that, by integrating multisource physiological signals such as functional near-infrared spectroscopy and surface electromyography with artificial intelligence algorithms, an intelligent perception framework for psychological and physiological indicators can enable real-time, accurate fatigue monitoring in adolescent athletes as well as personalized intervention [3]. Their study underscores the considerable potential of multimodal perception for applications in health science management. In extreme environments such as outer space, the multimodal visual perception for space robotic arms depends on the comprehensive integration of heterogeneous sensory data, including visible-light imaging, infrared sensing, depth cameras, and Light Detection and Ranging (LiDAR) [4]. This line of research is directed toward addressing major challenges in on-orbit servicing tasks, particularly target perception, trajectory planning, and compliant control, and has become one of the key directions for advancing the autonomous operational capabilities of space equipment. Taken together, recent advances suggest that multimodal perception technology is steadily moving beyond the constraints of individual modalities and application boundaries. Through continued innovation in sensor hardware integration and increasingly deep fusion with intelligent algorithms, it is enhancing the ability of complex systems to interpret dynamic environments and respond autonomously.

Although multimodal perception technologies have already found application in bionic robots, several challenges remain unresolved. Limited system integration often leads to bulky and inefficient architectures; existing algorithms still lack sufficient generalization capability in

dynamic, open-world scenarios; and the fusion of multisource heterogeneous information remains constrained in both efficiency and accuracy. At the same time, much of the existing literature focuses on particular sensing modalities or individual fusion approaches, while relatively few studies provide a comprehensive synthesis of the entire technical chain linking perception modalities, fusion strategies, and system-level challenges. Against this backdrop, the present study seeks to clarify both the developmental trajectory and the outstanding research issues in multimodal perception for intelligent bionic robots through a systematic review and analysis of its theoretical foundations, implementation approaches, fusion methods, and developmental bottlenecks. In doing so, it aims to offer practical directions for optimization as well as broader guidance for the development of efficient, robust, and adaptive embodied multimodal perception systems.

The remainder of this paper provides a systematic examination of embodied multimodal perception technologies for intelligent bionic robots. Its structure is organized as follows. Chapter 2 introduces the theoretical foundations of the field and outlines the implementation approaches and device-level basis of the principal perceptual modalities in intelligent bionic robots, including vision, audition, touch, and proprioception. Chapter 3 focuses on multimodal perceptual information fusion. It discusses the fundamental principles, strengths, application scenarios, and current research status of three representative fusion approaches: point cloud fusion, heterogeneous fusion, and bird's-eye-view (BEV) fusion. Chapter 4 analyzes the major challenges currently confronting these technologies, including limited device integration and inadequate generalization in static scenes. It also explores possible future solutions, particularly those based on integrated communication-sensing-transmission architectures, optoelectronic chips, and generative models. Chapter 5 concludes the paper by summarizing the main findings and highlighting their theoretical significance.

2 Theory and device implementation methods

2.1 Basic theory of multimodal perception

The theoretical foundations of multimodal perception can be traced to the information-processing mechanisms of biological systems, particularly those of humans. At its core, multimodal perception seeks to construct a more complete, robust, and accurate model of an environment or target by integrating heterogeneous sensor data drawn from different physical domains. The underlying premise is that any single perceptual channel is inherently limited in the information it can provide and is often susceptible to interference, whereas multisource information offers both complementarity and redundancy, thereby enhancing the fault tolerance, perceptual resolution, and scene-understanding capability of the overall system. Its theoretical basis spans several disciplines, including signal processing, information fusion, machine learning, and cognitive science. He et al pointed out that a theoretical framework grounded in sparse representation can address key problems in multimodal tactile information fusion [5]. By introducing shared sparse coding, this framework can uncover latent correlations and complementary structures across data from different modalities, thereby enabling robust recognition of environmental features such as material properties and providing both mathematical tools and algorithmic support for multimodal perception. In practical terms, multimodal perception theory guides the entire technical pipeline, from data

acquisition and feature extraction to spatiotemporal alignment and decision-level fusion. Tan established a relatively complete system of theories and methods for robot perception, including three-dimensional modeling of work scenes and vision-based localization [6]. In essence, these achievements are founded on collaborative perception models that combine vision with other sensing modalities, such as inertial and tactile sensing.

2.2 Perception modalities of intelligent bionic robots

Vision serves as the primary channel through which intelligent bionic robots acquire spatial information about their surroundings. Robotic visual perception has evolved well beyond monocular two-dimensional imaging and now encompasses a broader set of capabilities, including depth perception, motion estimation, and semantic understanding. The principal technical approaches include deep-learning methods based on monocular cameras, three-dimensional reconstruction using binocular or multicamera vision, and active depth sensing based on structured light or time-of-flight (TOF) technology. Ji et al reported that a construction polishing robot built on the YOLOv8 algorithm was able to detect wall marking points, and that data augmentation improved the robustness of the model in complex construction environments [7]. This represents a major approach to target detection and semantic segmentation based on monocular vision. Liu et al argued that a distance-measurement system based on binocular vision can recover depth information through stereo-matching algorithms and then combine it with two-dimensional image feature analysis to derive the three-dimensional spatial coordinates of a target [8]. In this way, it provides robots with accurate pose-perception capability and lays the foundation for non-contact operation.

Auditory perception provides intelligent bionic robots with the ability to recognize specific sound sources, interpret spoken commands, detect abnormal acoustic events, and localize sounds by processing audio signals. As such, it serves not only as an important channel for natural language human-robot interaction, but also as an effective complement to vision in environments where visibility is limited or where visual information is partially obstructed. A typical robotic auditory system includes a microphone array, an audio signal preprocessing unit, and machine-learning-based models for sound classification and speech recognition. Wang et al incorporated an auditory perception module that used deep learning to obtain natural-language operational commands and to perform audio recognition across twelve categories of sound signals [9]. By enabling task execution through audiovisual fusion, their study demonstrated that auditory perception, when integrated with vision, can improve a robot's understanding of both operational intent and environmental context through the combined processing of high-level semantic information carried by speech and low-level environmental cues conveyed by specific sounds.

Tactile perception enables robots to sense a wide range of physical interaction cues, including contact force, pressure distribution, texture, slip, and temperature. It is therefore fundamental to dexterous manipulation, safe human-robot interaction, and the stable grasping of soft and fragile objects. In terms of hardware realization, tactile sensing technologies have progressively evolved toward higher sensitivity, greater spatial density, increased flexibility, and tighter integration. A flexible grasping method and system for robots based on tactile sensors presents a method in which tactile sensors embedded in the robot, together with inertial measurement units, are used for object pose estimation on the basis of a pretrained model. This approach is crucial for adaptive grasping and constitutes a direct application of

tactile information in closed-loop control. Beyond conventional discrete force sensors, tactile sensing can also be implemented through fiber based devices and flexible conductive materials, such as liquid metals and carbon-nanotube composites [10].

Proprioception denotes a robot's capacity to sense its own internal state, including limb position, body posture, joint angles, velocity, and acceleration. It forms the foundation of motion control, trajectory planning, and coordinated execution. In robotic systems, proprioceptive information is typically obtained through encoders, inertial measurement units (IMUs), force sensors, torque sensors, and flexible bending sensors. A flexible grasping Method and System for Robots Based on Tactile Sensors notes that the system employs an IMU mounted on the robot to capture the pose of the end effector, a function that is essential to proprioceptive sensing and motion control. Encoders are used to measure joint angles, whereas IMUs provide acceleration and angular-velocity data for the trunk or limbs. Through sensor-fusion algorithms, the robot can estimate its body coordinate frame and joint-space states in real time. In addition, bending sensors embedded in flexible joints or artificial skin are able to continuously measure curvature, thereby supplying bionic robots with biologically inspired proprioceptive information and enabling more compliant and adaptive movement.

3 Fusion methods of multimodal perception information

3.1 Point clouds

Within multimodal perception systems, point cloud fusion is primarily used to process three-dimensional spatial information. It involves the handling, registration, feature extraction, and cross-modal fusion of discrete 3D point-set data acquired by sensors such as LiDAR. This approach can provide direct and highly accurate information about the three-dimensional geometry and distance of objects, while remaining largely unaffected by ambient lighting conditions. For this reason, point cloud fusion plays a vital role in tasks such as 3D object detection, mapping, and localization.

Its main areas of application are autonomous driving, robotic navigation, and three-dimensional reconstruction. Wu observed that, in point cloud processing, algorithms such as Point Pillars have been improved by introducing feature-enhancement modules during the transformation of point clouds into pseudo-images [11]. This design preserves more fine-grained information and, in turn, improves object-detection performance, reflecting continuing progress in point cloud processing itself. Despite these advances, point cloud fusion remains subject to several limitations. Raw point cloud data are inherently sparse and do not directly encode semantic attributes such as color and texture. Their performance can also deteriorate under adverse weather conditions, including rain, snow, and fog. Moreover, high-density point clouds impose substantial demands on computational power and storage capacity.

Recent work has progressed mainly in two directions. One line of research is devoted to the design of more efficient point cloud feature-learning networks so that both local and global features can be extracted more effectively. The other centers on deeper exploration of image-point cloud fusion, allowing the two modalities to complement each other at the data, feature, or decision level. In this context, the rich semantic cues available in images can be integrated with the accurate geometric information provided by point clouds, thereby improving the

performance of object classification, detection, and segmentation in complex scenes. As shown in Fig. 1, a point cloud fusion architecture based on virtual point convolution has been proposed to address the problem of excessive redundant computation caused by overly dense pseudo-point clouds generated from images [12].

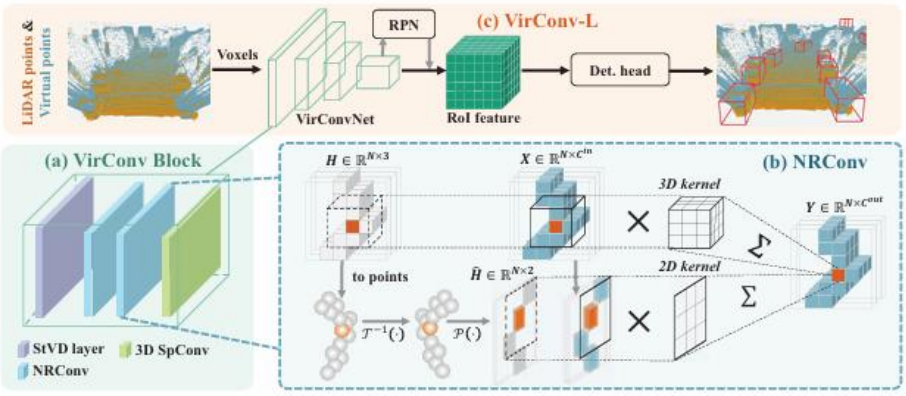

Fig. 1. Point cloud fusion architecture based on virtual point convolution technology [12].

3.2 Heterogeneous fusion

Heterogeneous fusion brings together sensor streams that differ fundamentally in their physical basis and statistical structure within a single perceptual framework. In robotic and intelligent sensing systems, this may involve combining visual inputs such as images and video with LiDAR point clouds, millimeter-wave radar signals, audio signals, tactile data, and even physiological measurements. The difficulty of this kind of fusion lies precisely in the heterogeneity of the inputs: the modalities differ sharply in dimensionality, spatiotemporal resolution, representational format, and information density. As a result, effective fusion depends on more than simply aggregating data. It also requires precise spatiotemporal synchronization and dependable cross-modal feature alignment.

The main strength of heterogeneous fusion is that it allows different sensing modalities to compensate for one another, reducing the blind spots and perceptual limitations associated with any single source of information. In robotics, usually use of visual data, IMU measurements, and wheel-encoder information for simultaneous localization and mapping (SLAM). Progress in heterogeneous fusion has been driven largely by advances in fusion strategy, especially the development of data-level fusion at an early stage, feature-level fusion at an intermediate stage, and decision-level fusion at a later stage. More recently, deep-learning-based end-to-end fusion networks have become increasingly common in robotic applications. These models are able to learn relationships across modalities automatically and generate unified joint representations with high informational richness, thereby enhancing perceptual performance in complex and dynamic environments. As shown in Fig. 2, a heterogeneous fusion framework based on multi-head attention is proposed to integrate multi-modal data with distinct dimensionality and resolution [13]. It unifies point-level and RoI-level fusion, combined with multi-task pseudo-LiDAR fusion, to achieve precise

spatiotemporal synchronization and feature alignment, thus enhancing object detection performance in complex scenes.

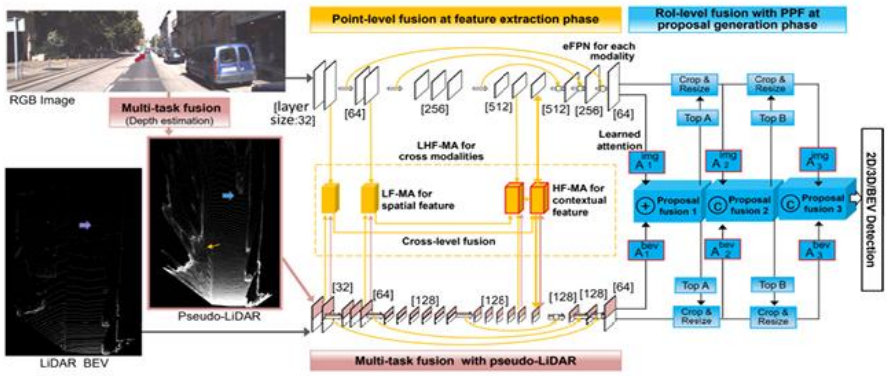

Fig. 2. Heterogeneous fusion framework based on multi-head attention [13].

3.3 BEV

BEV fusion has recently become one of the most important paradigms in multimodal perception, especially in autonomous driving. The core idea is to project sensor data captured from different viewpoints and modalities into a shared top-down two-dimensional coordinate space above the vehicle. This formulation offers clear advantages for perception and control. Because the resulting scene representation corresponds directly to the spatial layout used in vehicle planning and motion control, it is more intuitive for downstream tasks. It also mitigates the scale variation and occlusion effects introduced by perspective projection, while supporting functions such as drivable-area segmentation, object detection, and trajectory prediction. At present, BEV fusion is used mainly for global scene perception in autonomous vehicles. As illustrated in Fig. 3, Wei simplify cross-modal interaction through a unified BEV representation, employ self-attention mechanisms to model intermodal relationships, and incorporate a temporal fusion module to aggregate BEV features with spatiotemporal dependencies [14].

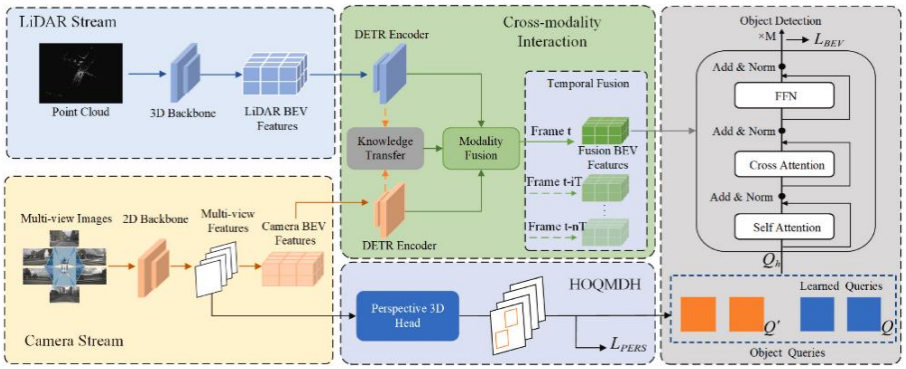

Fig. 3. BEV-based fusion framework [14].

4 Challenges and Future Prospects

4.1 Device Integration

Multimodal perception systems for intelligent bionic robots still face clear limitations at the hardware-integration level. In conventional system designs, sensing, communication, and computing modules are typically developed as separate units and interconnected through bus-based architectures. This often leads to bulky system layouts, high power consumption, and latency or bottlenecks in data transmission. Looking ahead, concepts borrowed from the communications domain suggest that integrated communication-sensing-transmission, or more broadly integrated communication-sensing-computing, will become a major direction for improving both device integration and overall system efficiency. This suggests that future perception hardware will move beyond the form of isolated sensors and instead evolve into integrated intelligent units capable of simultaneously performing environmental sensing, high-speed data transmission, and preliminary edge computing.

4.2 Static Scenes

Most existing deep-learning-based multimodal perception models depend heavily on large-scale, high-quality training datasets, yet such datasets are usually collected under constrained, relatively static, or otherwise structured conditions. As a consequence, these models often generalize poorly when deployed in truly dynamic, open, and unstructured environments. For instance, an object-recognition model trained under laboratory lighting and fixed environmental settings may suffer severe performance degradation outdoors, where illumination varies, weather conditions change, and unknown obstacles may appear. Furthermore, in real dynamic scenarios, spatiotemporal misalignment among data streams from different modalities can further intensify inconsistencies in the fusion process. Addressing weak generalization therefore requires models to learn feature representations that are more fundamental and robust, rather than merely fitting superficial statistical regularities in the training data. Achieving this goal will depend on advances in both algorithmic architecture and training paradigm.

4.3 Leveraging Integrated Optoelectronic Chips

In response to these challenges, two emerging technological directions deserve particular attention: integrated optoelectronic chips and generative artificial intelligence models. Integrated communication-sensing-computing systems for intelligent transportation demand integration at the hardware level. Within this framework, integrated optoelectronic computing chips could enable near-sensor computing by coupling optical-signal perception with electrical-signal processing directly at the sensing interface. From the algorithmic perspective, Artificial intelligence is central to equipping systems with intelligent perception and decision-making capabilities. Generative models, including diffusion models and large language models, can be used to synthesize large volumes of multimodal training data and to simulate rare or high-risk edge-case scenarios, thereby expanding available datasets and improving model generalization and robustness. At the same time, generative models can themselves function as powerful mechanisms for multimodal understanding and reasoning, directly strengthening robots' ability to interpret and respond to unknown environments. Taken

together, these two directions have the potential to drive a significant advance in intelligent bionic robot perception systems at both the hardware and algorithmic levels.

5 Conclusion

This paper presents a comprehensive review of the current research landscape and emerging trends in embodied multimodal perception technologies for intelligent bionic robots. First, the technological trajectory from single-modality sensing to multimodal information fusion is systematically traced, highlighting the embodied-intelligence-driven perception-action loop as a prevailing direction in the evolution of intelligent robotic systems. Then, the bionic foundations underlying multimodal perception, the implementation principles and hardware development of major perceptual modalities is disscussed, including vision, audition, touch, and proprioception. On this basis, the paper further investigates three representative paradigms for multimodal information processing, namely point cloud fusion, heterogeneous fusion, and BEV spatial fusion, with particular attention to their applications and technical characteristics. The analysis also indicates that existing technologies continue to face significant challenges, most notably limited device integration and insufficient generalization in dynamic environments. Future progress in this area will depend on a coordinated software-hardware framework that integrates communication-sensing-transmission architectures, integrated optoelectronic chips, and generative artificial intelligence models. Through the systematic synthesis and analysis of the full technical chain in this field, providing both a clearer theoretical reference and practical directions for the development of intelligent bionic robot perception systems with greater efficiency, adaptability, and generalizability.

References

[1] Dong, J. C., Zhou, L., Sun, Q. Y., et al.: Bionic intelligent robot technology: from morphology to embodied intelligence. Information and Control. pp. 1-19 (2026).

[2] Kong, W. B., Chen, X. J.: Research on fully autonomous inspection technology of robot dog in new energy intelligent power station based on multimodal fusion perception technology. Wind Energy. pp. 76-85 (2025).

[3] Lin, Y. T., Dong, X. X.: Multimodal intelligent perception and AI fusion for fatigue monitoring of adolescent athletes: technical framework and application prospects. Youth Sports. pp. 66-71 (2025).

[4] Hu, Y. H., Wu, L. G., Chen, J. G., et al.: A review of multimodal visual perception and manipulation technologies for space manipulators. Journal of Harbin Institute of Technology. Vol. 57, no. 12, pp. 1-21 (2025).

[5] He, K. F.: Research on theory and application of robot tactile perception based on sparse representation. Nanchang University, Nanchang (2021).

[6] Tan, M.: Intelligent Robot Perception and Control. Institute of Automation, Chinese Academy of Sciences, Beijing (2020).

[7] Ji, Y. J., Zheng, H., Jin, Y. P., et al.: Research on vision system of building grinding robot. Construction Technology (Chinese & English). Vol. 55, no. 3, pp. 44-53 (2026).

[8] Liu, D. R., Han, C. H.: Research on robot binocular ranging system based on machine vision. Electrical Industry. pp. 16-21 (2026).

[9] Wang, Y. F., Ge, Q. B., Liu, H. P., et al.: Perception operating system based on fusion of robot vision and hearing. Journal of Intelligent Systems. Vol. 18, no. 2, pp. 381-389 (2023).
[10] Li, C X., Liu, Y Q., Han, D D. et al.: A tactile gripper on an optical fiber for perception and actuation. Nature Communication Vol. 16, pp. 11513 (2025).
[11] Wu, X. P., Peng, L., Yang, H. H., Xie, L., Huang, C. X., Deng, C. Q.: Sparse fuse dense: Towards high quality 3D detection with depth completion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. pp. 5408-5417.
[12] Wu, H., Wen, C.L., Shi, S.S., Li, X., Wang, C.: Virtual sparse convolution for multimodal 3D object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. pp. 21653-21662.
[13] 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 Transactions on Intelligent Transportation Systems. Vol. 24, no. 2, pp. 1692-1701 (2023).
[14] Wei, M.: BEV-CFKT: A LiDAR-camera cross-modality-interaction fusion and knowledge transfer framework with transformer for BEV 3D object detection. Neurocomputing. Vol. 582, pp. 127527 (2023).