Deep Learning for Multi-Modal Sensor Fusion in Robotics: A Comprehensive Survey of Perception and Navigation Systems

[1]Vikash Kumar Verma, [2]Dr. Sourabh Mandaloi, [3]Ruchi Chaturvedi
[1]M. Tech Scholar, Department of Computer Science, SAM College, Bhopal, india
[2]Associate Professor, Department of Computer Science, SAM College, Bhopal, india
[3]Assistant Professor, Department of Computer Science, SAM College, Bhopal, india
[1]Jecvikas82@yahoo.com

Abstract: *A comprehensive survey is presented on the recent progress in deep learning multi-modal sensor fusion for robotic perception and navigation. As autonomous systems are increasingly operating in complex unstructured environments, having multiple sensing modalities such as RGB cameras, LiDAR, IMUs, GPS, etc. has become imperative to remedy the limitation of any one sensor. Deep learning has further enhanced this integration by facilitating strong, scalable, and adaptive fusion of heterogeneous data streams. Architectures such as CNNs, RNNs, Transformers, and GNNs are being used for feature extraction and fusion towards situational awareness and decision-making. The paper categorizes the fusion techniques into three levels: early, intermediate, and late fusion, analyzing the pros and cons of each. Sensor calibration, temporal synchronization, noise, and computational feasibility for real-time realizations are outlined among the primary challenges. This review emphasizes how deep learning assists not only in automatic feature extraction but also in engendering context-aware, resilient robot behavior in dynamic environments. The paper concludes that deep learning-based multi-modal fusion will form a critical backbone enabling future intelligent robotic systems to operate reliably and autonomously in myriad scenarios.*

Keywords: Multi-modal sensor fusion, Deep learning, Robotic perception, Autonomous navigation, CNN, RNN, Transformer, Graph Neural Networks (GNNs), Sensor integration.

1. Introduction

Advanced machine learning, especially deep learning, has empowered the fusion with autonomous systems to enrich advances in robotics in these recent years. A fundamental challenge with robots of late operating in complex and dynamic real-world environments is not just to perceive the world accurately but to interpret and act upon it in time. Conventional sensor systems, such as cameras or LiDAR, provide very partial information, leading to interpretations of the environment that are usually ambiguous or incomplete. This calls for what is known as multi-modal sensor fusion [1]. Sensor fusion is a set of methods that combine data from different sensing modalities to provide a more enriched, consistent, and reliable representation of the environment. The idea is to improve a robot's perception, localization, and navigation, hence allowing it to safely and efficiently perform in various tasks, be it autonomous driving, drone navigation, or industrial automation. Deep learning for multi-modal sensor fusion is, in essence, aimed at surpassing the limitations that arise from using just a single sensor system, enabling robots to independently navigate and make decisions in a complex environment [2].

Multi-modal sensor fusion has become the backbone of modern robots since it allows for assembling sensors that gather complementary information from an environment and merging such information to enhance the accuracy, strength, and versatility of a robotic system. The strength of one kind of sensor poses its limitations on another: viz., vision-based sensors are strongly affected by illumination; lidar provides the most precise depth information but may suffer from some surface properties; IMUs may drift [3]. Fusing such diverse information sources provides a better, more reliable understanding of the ring. This fusion becomes necessary where single-modality perception cannot work, e.g., low-light or cluttered environments. As a consequence, robots in multi-modal systems can respond to a more diversified set of objects and obstacles, improve their spatial awareness, and enhance their decision-making ability for more autonomous and safe tasks in navigation and task execution [4].

Arguably, the most suitable offering of deep learning to robotics is the automatic real-time analysis of huge amounts of sensor data in multi-modal fusion applications. The classical algorithms depended on handcrafted features and rule-based systems that were rather rigid when confronted with the intricacies and variations of the real-world environment. On the other hand, deep learning algorithms, primarily CNNs, RNNs, and more recently transformer models, automatically learn discriminative features from the encoded raw sensor data with respect to the desired perception tasks, rendering perception systems far more accurate and adaptive [5]. These deep models shine in the presence of multi-sensory data, barring complex high-dimensionalities-from RGB images to depth maps and point clouds-for assessment purposes. By bringing in this technology, robots can achieve a far greater degree of autonomy, equipping themselves with perception, prediction, and navigation skills all far less dependent on human intervention [6]. Having this ability to generalize given large datasets and to quickly adapt itself to different environments makes deep learning paramount in the making of autonomous robots, which are now increasingly called to operate in truly unstructured, and unpredictable environments-urban streets, industrial sites, or even disaster areas.

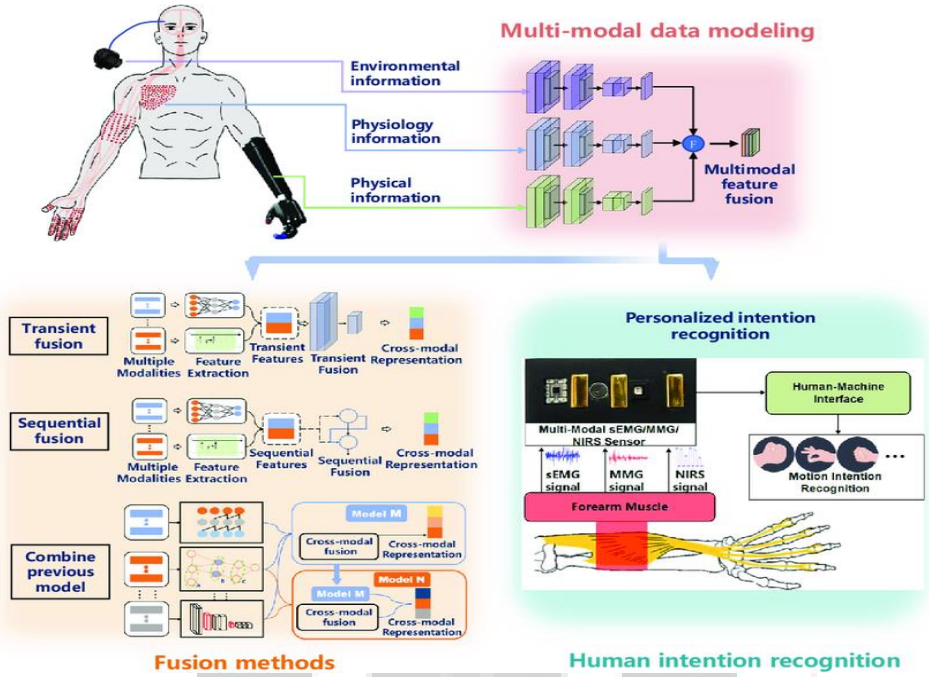

Figure 1 Multi-modal Fusion for Personalized Human Intention Recognition [7]

2. Fundamentals of Multi-Modal Sensor Fusion in Robotics

Multi-modal sensor fusion in robotics is a process that integrates information from different sensor types ranging from RGB cameras, LiDARs, IMUs, and GPS to ultimately equip the robot with perception and decision-making abilities. By integrating the strengths of different sensors, robots can minimize their respective weaknesses, improving overall accuracy and robustness in dynamic environments. The fusion itself can happen at several stages: Early fusion considers the fusion of data at the raw level, intermediate fusion at the feature level, and late fusion at the decision or output level [8]. Traditional methods have mostly relied on rules, such as Kalman filters, and even with explicit definition of features, criteria for fusion, and time constraints, deep learning enables automatic extraction of features and fusion, offering a more versatile and adaptive way of fusion. Still, issues such as calibration, noise, and synchronization pose hard challenges with computational constraints; thus, designing an optimized fusion system is vital in real-time robotic applications [9].

A. Types of Sensors Used in Robotics

Various sensors allow a robot to be aware of its environment: each comes with its own pros and cons. In general, RGB cameras are used for object recognition and scene understanding but cannot be relied upon in low-light conditions. LiDAR measures distances accurately for mapping and obstacle detection in tough situations such as darkness or fog; nonetheless, in cluttered environments, it gets limited in terms of range and resolution. IMUs track movements and help in stabilizing robots, especially when operating in GPS-denied environments; however, drift errors become an issue as time goes on. GPS sensors are used in outdoor navigation with the best positional accuracy but are disrupted when used in obstructed regions, such as with high-rise buildings or indoors. Audio sensors, mainly microphones, assist in localizing sounds and recognizing speech; however, one needs to use noise-filtering techniques, or else they may fail in noisy surroundings. By fusing these sensors, one can overcome the limitations of the individual sensors and have a richer understanding of the environment.

1 RGB cameras

RGB cameras are the most common across all types of robots sensing environment for dark color images. They work quite similarly to human vision, providing the visual data necessary to execute tasks such as recognizing objects, understanding scenes, or visual navigation. RGB cameras help with the identification of objects, tracking movement, and interfacing with the environment. These cameras do have their drawbacks. These cameras are susceptible to lighting conditions, which is to say that their effectiveness diminishes when in low light roles in too much light [10]. Also, considerations of shadow, reflection, and poor illumination distract them, degrading their reliability under various conditions. Yet, despite these issues, they continue into the mainstream in almost every robotic vision system because they produce rich visual information [11].

2 LiDAR (Light Detection and Ranging)

A LiDAR is a laser-based distance-measuring device that emits light and measures the time it takes for the light to be reflected back from objects in the environment. With the capability to generate accurate 3D maps with large spatial resolution, these systems prove to be a boon in obstacle detection, path planning, and environment modelling [12]. In navigation, LiDAR sensors are also handy as they are able to measure distances in conditions wherein optical systems such as cameras might fail-like in darkness or fog. Despite the usefulness, LiDAR systems

show some drawbacks. They could have a limited range and resolution in highly structed or cluttered environments, thus causing problems when detecting smaller objects or objects farther away [13].

3 IMU (Inertial Measurement Unit)

An inertial measurement unit (IMU) is a type of sensor used for the measurement of movement parameters, such as acceleration, angular velocity, and orientation. IMUs are vital to robotics in motion tracking, position estimation, and stabilizing the robot through motion. They have the best application where GPS signals are weak to almost non-existent, such an environment that might be indoors or underground [14]. IMUs effectively let the robot operate with motion while confronting these types of environments by giving out real-time data concerning their motion. The downside is that the IMUs can have drift errors in their measurements as time goes by; this one mean that the IMU estimates of position and orientation become less useful and inaccurate with time without an external measure. This is why they are often used in fusion systems alongside others [15].

4 GPS Sensors

Global Positioning System (GPS) sensors find extensive use in outdoor robotic applications for the computation of exact geometric location data. Receiving signals from a constellation of satellites allows robots to pinpoint their position with great precision over large spatial areas [16]. Thus, autonomous vehicles and drones, as well as any robot working in the large-sized outdoors where precise navigation gets employed by the term, finds GPS highly useful for route planning, geofencing, or outdoor navigation. With all its potentials, GPS sensors, however, do have some demerits. The tools tend to do well depending on the direct view of satellites and tend to limit themselves as soon as they become obstructed in some so dense urban environment, indoor location, or under heavy tree cover from the reflection or blocking of signal [17].

5 Audio Sensors

Audio sensors convert sound waves into a proper form, usually using microphones. For example, audio signals are used by robots that facilitate sound localization, recognize speech, and are environment-aware. Likewise, search-and-rescue robots employ audio sensors to pick up on human voices or any other sounds that may imply the presence of people or hazards [18]. Human-robot interaction also depends on microphones for the recognition of voice commands. Nevertheless, an inferior-quality audio sensor can pose a problem: it captures background noise and may have trouble differentiating between the noises of concern and ambient ones, especially in louder environments. Hence, they tend to be less useful in some environments unless in conjunction with the noise-mitigation techniques [19].

B. Fusion Levels

Sensor fusion is simply the joining of data generated by multiple sensors in a bid to provide an improved comprehension of the environment by the robot. Fusion in this case can occur at different levels of the processing pipeline, generally divided into early fusion, intermediate fusion, and late fusion. Early fusion refers to the combination of raw data coming from different sensors before individual analysis is performed. For instance, one could combine raw images coming from a camera with depth information coming from LiDAR at the pixel level to generate a richer representation of the scene [20]. While early fusion attempts to leverage fully the complementary nature of different sensors, it usually places stringent requirements on synchronization or alignment of data, rendering it computationally expensive and complex. Intermediate fusion is the stage where sensor data are individually processed to extract useful features (such as object contours or motion patterns) that are then combined for further processing [21]. This method works better in terms of computing power than early fusion and has more flexibility in terms of data treatments, thus making it more adaptable to different implementations on robots. Late fusion is the most modular approach in which data from each sensor is processed separately until the decision is made when the fusion of the output of each one occurs. For example, separate object detectors on RGB data and LiDAR data output their predictions that are subsequently fused to make the final decision on what actions the robot takes. Late fusion is computationally less costly and easier to design but cannot exploit the interrelationships across different sensor modalities to their full extent, resulting in cases when performance could be less desirable [22].

C. Challenges in Multi-Modal Sensor Fusion

Although multi-modal sensor fusion in robotics is highly consistent with its common purpose of working together, some major inconsistencies have to be resolved towards successful implementation. The major concerns include sensor calibration, which pertains to the alignment of the sensor data on the spatial-temporal spectrum. For instance, a LiDAR sensor may have a far different field of view and resolution compared to the RGB camera, making it challenging to combine data from them without introducing errors [23]. Sensor noise is another one, and these noise levels vary with respect to the environment, so a specific data set might contain noise or be unreliable. For example, an IMU sensor may drift, and a LiDAR might run into reflective surfaces distorting depth measurements. Synchronizing data across sensors is another hindrance because these sensors subscribe to different sampling rates. For example, a camera may be designing a picture every 30 seconds, whereas a LiDAR can be analysing his scene texture at a very low number of frames per second. Temporal alignment of the data from both sensors prior to fusion is mandatory for meaningful fusion results [24]. Real-time feasibility is another bottleneck. Robotics-based applications in dynamic environments pitch immediate decisions over fused sensor data. Large

volumes of multi-modal data throughputs are crucial to ensure real-time performances, which in itself is a bottleneck of computational capacities. Then comes the question of robustness, for the ability of a robot to carry out its tasks in unstructured environments is paramount. The fusion has to adapt where either of the sensors fails or gets obstructed, or faces some environmental issues like poor lighting or dense fog [25].

3. Traditional vs. Deep Learning-Based Approaches

Conventional sensor fusion techniques in robotics were mostly implemented with handcrafted features and rule-based algorithms. The practitioner manually selects relevant characteristics from sensor data: such features could be edges from images or distances from LiDAR. Then an algorithm, such as a Kalman filter or a particle filter, will combine the information to provide an estimate. These traditional techniques had served well under controlled environments with predictable data but struggled when dealing with noisy, incomplete, or complex realities of real-world data [26]. Also, the techniques needed heavy weight domain knowledge and constant tuning, which restricted their flexibility and adaptability. On the other hand, deep learning methods have energized the field of sensor fusion by automating the extraction of relevant features and end-to-end sensor data fusion. Different deep learning models, such as convolutional neural networks (CNNs) for image data and recurrent neural networks (RNNs) for temporal data, can learn complex correlations existing in multi-modal sensor data without feature engineering. Such models work well with noisy and unstructured data and exhibit good generalization ability in varying environments [27]. Another significant advantage of deep learning lies in its capability to learn from big data and robustly handle diverse sensor modalities so that every environmental variation can be considered. Nevertheless, fusion methods derived from deep learning have issues that are challenging to overcome, including the requirement of lots of annotated data for the training, large computational costs, and hard-to-understand model decisions. Meanwhile, the deep learning paradigm gives some flexibility, accuracy, and scalability that traditional methods hard to achieve-most of the time-end up being the default in most systems [28].

4. Deep Learning Techniques for Multi-Modal Fusion

Deep learning techniques in multi-modal fusion use state-of-the-art neural network architectures for data integration from various sensors including cameras, LiDARs, IMUs, and GPSs to enhance perception and decision-making in robotics. CNNs process images for spatial feature extraction; RNNs and LSTM handle temporal analysis that robots use for tracking movement and predicting future states. Transformer models are really good at catching long-range dependencies from heterogeneous sensor inputs, while GNNs suit sensor data that has a relational structure. Attention mechanisms dynamically select relevant sensor information, boosting the performance and efficiency of fusion steps. These deep learning techniques enable robots in understanding and navigating complex environments by automatically discovering the most appropriate way for combining multi-modal data.

A. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are among the most extensively used deep learning techniques in computer vision and multi-modal sensor fusion. These networks were discerned to process grid-like data such as images; thus, a series of convolutional layers is applied to an input image to learn spatial hierarchies of features automatically. In the realm of multi-modal fusion, CNNs find a great use in combining visual data with corresponding depth data acquired from LiDARs or other spatial sensors [29]. CNNs have it in their architecture to learn features and objects at different levels of abstraction and thus are able to effectively fuse the different sensor modalities, thereby improving tasks such as robot-based object detection, segmentation, and localization. These networks can handle sensor data of dimensionality and are often used as the foundational layer for deeper and more complex multi-modal fusion architectures [30].

B. Recurrent Neural Networks (RNNs) and LSTMs

The random neural networks and long short-term memory networks (LSTMs) are developed to capture and work upon sequential data, hence when working with robotics-based applications where time dependency is dictated by sensor data, such data will include outputs from an IMU, GPS, or auditory sensors. RNNs and LSTMs model temporal relationships well by exploiting the nature of hidden states that can remember the past for long so as to grasp long-term dependence [31]. For the multi-modal sensor fusion, one would employ RNNs and LSTMs to process time-series data describing movement paths of a robot or predicting future states on the basis of sensor information received over time. This allows RNNs and LSTMs to be applied in such fields as visual odometry, motion tracking, and adaptation to a dynamically changing environment, in which time context and sequence prediction are very much needed [32].

C. Transformer-Based Models

Transformer models were first introduced to NLP applications, but they are now gaining popularity in multi-modal fusion tasks owing to their handling of long-range dependencies and efficient capture of global concepts. Transformer's mechanism of self-attention allows dynamic weighing of different parts of the input, according to their importance, and hence is very useful in fusing heterogeneous data from different sensor types [34]. For instance, while fusing images with LiDAR point cloud data, the transformer model will effectively learn and combine which features are the relevant ones from each modality. Due to their parallelizability and scalability,

transformers find use in demanding and complicated multi-modal tasks of real-time applications in robotics, such as tracking, scene understanding, and multi-sensor localization [35].

D. Graph Neural Networks (GNNs)

Graph Neural Networks (GNNs) are an emergent category for models used in multi-modal sensor fusion in robotics, especially when data comes with an inherent graphical structure. Within robotics, GNNs find use cases in multi-modal data fusion for sensors that convey relational or spatially distributed information, such as LiDAR point clouds, robot states, or objects within an environment [36]. A GNN system can learn to propagate information through the nodes of a graph to systematize spatial and topological dependencies across different elements (objects, locations, or events). Hence, for multi-modal fusion, GNNs are very apt in fusing sensor data represented heterogeneously so that the robot can perform tasks like navigation in dynamic environments, multi-object tracking, and environment mapping [37].

E. Attention Mechanisms in Fusion

Attention mechanisms are considered a distinguished feature of deep learning models for enabling them to concentrate on the crucial parts of input data with respect to a given task. Multi-modal sensor fusion views attention mechanisms as means for such models to choose dynamically which sensor inputs are relevant to the task at hand and to weigh their contributions from different modalities with respect to context [38]. Attention mechanisms could, for example, let a robot prioritize LiDAR-based information while detecting obstacles and visual data while recognizing objects when navigating through a cluttered environment. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers are usually accompanied by attention mechanisms to help the fusion process become more interpretable and better performing under noises, incomplete information, and ambiguous sensor inputs. This facility is equally valuable in decision-making, path planning, and multi-modal perception in robotics [39]. The figure 2 shows the multi-modal fusion framework, where image and point cloud features are jointly processed through patch embedding, transformer-based cross attention, and convolutional fusion for enhanced feature representation.

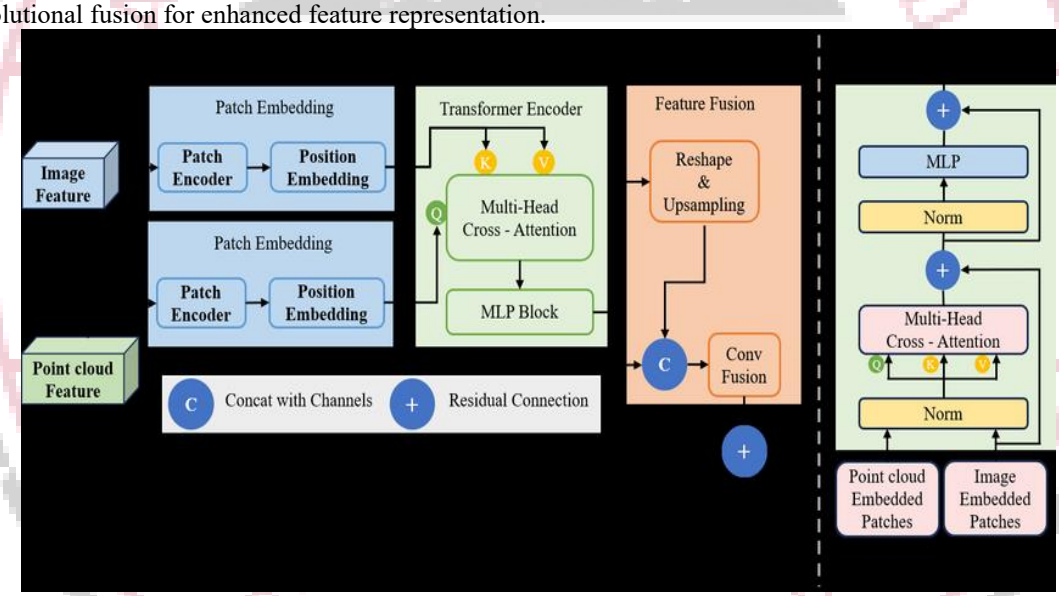

Figure 2 (a) The structure of the ViT multimodal fusion module, and (b) the Transformer Encoder. The fusion module consists of three sequential steps: patch embedding, Transformer encoder and feature fusion. The main blocks of Transformer encoder are MultiHead Cross-Attention and MLP blocks [40]

5. Multi-Modal Fusion for Robot Perception

The concept of multi-modal fusion for robot perception is a fundamental one in robotics, through which the robots aim to obtain a stronger and more complete understanding of their surroundings using various types of sensor data. One sensor modality includes an RGB camera, LiDAR, IMU, and radar [41]. Each sensor has its pros and cons; no single sensor can give perception with complete information. Cameras often provide detailed information but lack sufficient features to work in dimlight or when obstacles appear in between; on the other hand, LiDAR provides accurate depth measurements but does not give the texture or color information that cameras offer [42]. Hence, by fusing all the different types of sensor data, robots and their surroundings can overcome limitations introduced by the use of an individual sensor, granting them better awareness of their surroundings. Multi-modal fusion allows robots to more reliably detect and understand objects, track motion, and navigate complex environments [43]. All sensors complement one another by providing information that one would lack if implemented.

The fusion process integrates data at different levels: early, intermediate, or late, with advanced machine learning models, in particular deep learning, facilitating this process. These models learn to fuse sensor data in a manner

that optimizes perception and maximizes decision-making capabilities of the robot. For instance, CNNs can process visual data, whereas RNNs can interpret time-series data from an IMU [44]. By being trained on enormous datasets, deep learning approaches are able to fairly accurately detect objects and avoid obstacles in abnormal and dynamic settings or to grasp semantic knowledge in more difficult scenarios. Moreover, with the assistance of attention mechanisms, robots give precedence to the most relevant data available from their sensors. Multi-modal fusion plays a core role in assisting robots to autonomously and efficiently operate and adapt to a wide range of real-world scenarios and environments [45].

6. Challenges

Multi-modal sensor fusion in robotics involves several challenges that must be resolved to guarantee reliability and efficiency throughout performance. Calibration and alignment of sensors: distinct sensors may have different fields of view, resolutions, and sampling rates, causing data alignment to be inaccurate. More problems come in because of noise and reliability issues, as each sensor modality has faults like cameras having sensitivity in low light or IMUs drifting. Synchronization of collected data also poses another challenge since these sensors operate differently with respect to their sampling rate, and improper synchronization will yield inappropriate results from fusion [46]. Another aspect of concern is that it must be done in real-time as required by dynamic setups. Hence, with the appropriate parameters, computational effort ought to be weighed down with that of considerations related to efficiency when dealing with such sizable data. Environment adaptability to accommodate the robot is imperative; for instance, variations in illumination or obstacles are normal. Scalability and adaptability to supposedly include new sensors and operate in a new environment should be additional considerations. Moreover, providing an intelligent interpretation of complicated multi-sensory data to produce intermediate inferences for applications such as object detection, recognition, and navigation is still a challenge and needs more advanced techniques like deep learning for feature extraction and fusion. Being able to tackle such issues is very vital to let the robot work autonomously and efficiently in a myriad of real-world scenarios [47].

With extensive advances in the field of autonomous vehicles and systems, they have attained improvements for mobility and safety through mechanized decision-making frameworks [48], [49]. Yet, the existing techniques generally cannot operate well under complex settings or sufficiently embed interactions with surrounding vehicles. To take care of these shortcomings, AUTO framework combines deep reinforcement learning with multi-modal perception for adaptable decision-making in diverse environments, employing graph-based methods for state representation and parameterized action structures for lane following versus lane changing decisions [50], [51]. Additionally, there exist dynamic obstacle avoidance methods that perfectly suit real-time challenges in dynamic environments, with neuromorphic vision sensors feeding models designed through paradigms of deep reinforcement learning [52], [53]. Other avenues of advanced multi-modal learning, including those of the Uni-Modal Teacher, enrich the learning of modality-specific representations for problems such as modality failure and thereby enhance downstream multi-modal task learning [54], [55]. Furthermore, multi-modal tactile sensing and improved 3D object detection using point clouds and RGB images give significant leverage in texture recognition and object localization [56]. Lastly, which have been developed recently, foster another method in trajectory estimation via Graph Neural Networks (GNNs) to better aid robot navigation through complex outdoor environments, hence showing great development in perception and decision-making for autonomous systems [57].

Table 1 Comparative Analysis of Multi-Modal Fusion Approaches in Robotics and Autonomous Systems

Reference	Main Focus	Key Methods	Primary Goal	Results/Outcomes
[48]	Autonomous driving decision-making framework with deep reinforcement learning and multi-modal perception.	Deep reinforcement learning (DRL), graph-based model, hybrid reward function.	Optimize decision-making and vehicle actions for improved safety, traffic efficiency, and passenger comfort.	State-of-the-art performance in macroscopic and microscopic autonomous driving tasks.
[49]	Dynamic obstacle avoidance using a hybrid DRL-based multi-modal sensory approach.	Event camera, deep reinforcement learning, spiking neural network, unsupervised representation learning.	Enhance dynamic obstacle avoidance with a hybrid DRL-based multi-modal sensory approach.	Outperforms existing dynamic obstacle avoidance methods, especially for moving obstacles.

[50]	Improving multi-modal fusion through a novel approach that resolves modality failure.	Uni-Modal Teacher framework combining fusion objectives and uni-modal distillation.	Solve modality failure in multi-modal fusion to improve individual modality representations.	Improved representation learning and multi-modal task performance with a significant boost in accuracy.
[51]	Multi-modal bionic finger tactile sensor for texture recognition with wavelet transform and CNN.	Multi-modal bionic finger tactile sensor, wavelet transform, CBAM-CNN for feature fusion.	Improve tactile texture recognition by fusing multi-modal tactile signals and CNN-based models.	Achieved high tactile texture recognition accuracy across multiple datasets with CBAM-CNN.
[52]	Improved anchor generation for 3D object detection by using 2D guidance and multi-layer fusion.	2D detector-based anchor generation, multi-layer fusion model with BEV representation for point cloud.	Optimize 3D object detection by using guided anchor generation and multi-layer feature fusion.	Improved anchor generation for 3D object detection with better precision and performance on KITTI.
[53]	Analysis of perception fusion driving in autonomous driving systems.	Perception fusion techniques applied in autonomous driving with AI-driven decision systems.	Evaluate and enhance autonomous driving perception systems with AI-driven decision systems.	Perception fusion analysis for autonomous driving, highlighting the benefits of multi-source sensor fusion.
[54]	Open-source design for a multi-modal tactile sensing module for robotic hands.	Compliant tactile sensing module design with 3D printed molds, ROS support.	Design and fabricate an open-source, compliant multi-modal tactile sensor for robotic applications.	Effective tactile sensing module for robotics, with easy assembly and wide applicability in various robots.
[55]	Bi-stage multi-modal fusion method for high-precision 3D instance segmentation in workshops.	RGB-D multi-modal fusion, 2D prior information, correlation filtering for 3D segmentation.	Enable high-precision 3D instance segmentation without 3D labels using multi-modal fusion.	Achieved accurate 3D segmentation in a production workshop with improved performance over RGB-only methods.
[56]	Trajectory prediction using multi-modal sensory inputs (RGB, LiDAR, odometry) for robot navigation.	Graph Neural Networks (GNN), attention-based model for trajectory success probability prediction.	Improve trajectory prediction and robot navigation in complex environments with multi-modal fusion.	Improved trajectory prediction with increased navigation success rate and decreased false positives.
[57]	Learning manipulation tasks through video-captioning and multi-modal fusion for robot task execution.	Multi-modal fusion for video captioning, action classifier, keyframe alignment, and command decoder.	Train robots to learn manipulation tasks from human demonstrations using multi-modal fusion.	Significant improvements in translation accuracy of commands for manipulation tasks in robots.

7. Conclusion

Deep learning-fortified multi-modal fusion constitutes a new paradigm disruptor in robotics perception and autonomy. Considered truly complementary, the sensor combinations lend additional situational awareness, navigation, and decision-making capabilities to the robots. While classical methods were restricted by human-made rules and hard designs, deep learning-based models could adopt a flexible data-driven approach to extract and fuse rich and pertinent features with complex, noisy, and asynchronous sensor data streams. Despite the hurdles in sensor misalignment and high computational overhead, the ever-growing developments of advanced

neural architectures and fusion methods, such as those based on attention and transformer models, suggest that scalable and near real-time implementations may be realized soon in dynamic scenarios. Multi-modal fusion is, thus, firmly set to continue driving the agenda for the generation of truly robust, highly adaptive, and highly intelligent robotic systems into the coming years.

References

[1] Bednarek, M., Kicki, P., & Walas, K. (2020). On robustness of multi-modal fusion—Robotics perspective. *Electronics*, *9*(7), 1152. https://doi.org/10.3390/electronics9071152
[2] Huang, K., Shi, B., Li, X., Li, X., Huang, S., & Li, Y. (2022). Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*. https://doi.org/10.48550/arXiv.2202.02703
[3] Han, X., Chen, S., Fu, Z., Feng, Z., Fan, L., An, D., ... & Xu, S. (2025). Multimodal fusion and vision-language models: A survey for robot vision. *arXiv preprint arXiv:2504.02477*. https://doi.org/10.48550/arXiv.2504.02477
[4] Li, Z., Zhou, A., Pu, J., & Yu, J. (2021). Multi-modal neural feature fusion for automatic driving through perception-aware path planning. *IEEE Access*, *9*, 142782-142794. https://doi.org/10.1109/ACCESS.2021.3120720
[5] Tang, Q., Liang, J., & Zhu, F. (2023). A comparative review on multi-modal sensors fusion based on deep learning. *Signal Processing*, *213*, 109165. https://doi.org/10.1016/j.sigpro.2023.109165
[6] Zeng, H., & Luo, J. (2022). Construction of multi-modal perception model of communicative robot in non-structural cyber physical system environment based on optimized BT-SVM model. *Computer Communications*, *181*, 182-191. https://doi.org/10.1016/j.comcom.2021.10.019
[7] Xia, Haisheng & Zhang, Yuchong & Rajabi, Nona & Taleb, Farzaneh & Yang, Qunting & Kragic, Danica & Li, Zhijun. (2024). Shaping high-performance wearable robots for human motor and sensory reconstruction and enhancement. Nature Communications. 15. 10.1038/s41467-024-46249-0.
[8] Mohd, T. K., Nguyen, N., & Javaid, A. Y. (2022). Multi-modal data fusion in enhancing human-machine interaction for robotic applications: a survey. *arXiv preprint arXiv:2202.07732*. https://doi.org/10.48550/arXiv.2202.07732
[9] Qiu, Z., Martínez-Sánchez, J., Arias-Sánchez, P., & Rashdi, R. (2023). External multi-modal imaging sensor calibration for sensor fusion: A review. *Information Fusion*, *97*, 101806. https://doi.org/10.1016/j.inffus.2023.101806
[10] Lee, J., & Ahn, B. (2020). Real-time human action recognition with a low-cost RGB camera and mobile robot platform. *Sensors*, *20*(10), 2886. https://doi.org/10.3390/s20102886
[11] Skoczeń, M., Ochman, M., Spyra, K., Nikodem, M., Krata, D., Panek, M., & Pawłowski, A. (2021). Obstacle detection system for agricultural mobile robot application using RGB-D cameras. *Sensors*, *21*(16), 5292. https://doi.org/10.3390/s21165292
[12] Lopac, N., Jurdana, I., Brnelić, A., & Krljan, T. (2022). Application of laser systems for detection and ranging in the modern road transportation and maritime sector. *Sensors*, *22*(16), 5946. https://doi.org/10.3390/s22165946
[13] Jiang, A., & Ahamed, T. (2023). Navigation of an autonomous spraying robot for orchard operations using LiDAR for tree trunk detection. *Sensors*, *23*(10), 4808. https://doi.org/10.3390/s23104808
[14] Samatas, G. G., & Pachidis, T. P. (2022). Inertial measurement units (imus) in mobile robots over the last five years: A review. *Designs*, *6*(1), 17. https://doi.org/10.3390/designs6010017
[15] Semwal, V.B., Gaud, N., Lalwani, P. *et al.* Pattern identification of different human joints for different human walking styles using inertial measurement unit (IMU) sensor. *Artif Intell Rev* **55**, 1149–1169 (2022). https://doi.org/10.1007/s10462-021-09979-x
[16] Weiss, P. (2021). The Global Positioning System (GPS): Creating satellite beacons in space, engineers transformed daily life on earth. *Engineering*, *7*(3), 290-303. https://doi.org/10.1016/j.eng.2021.02.001
[17] Gharajeh, M. S., & Jond, H. B. (2020). Hybrid global positioning system-adaptive neuro-fuzzy inference system based autonomous mobile robot navigation. *Robotics and Autonomous Systems*, *134*, 103669. https://doi.org/10.1016/j.robot.2020.103669
[18] Albustanji, R. N., Elmanaseer, S., & Alkhatib, A. A. (2023). Robotics: five senses plus one—an overview. *Robotics*, *12*(3), 68. https://doi.org/10.3390/robotics12030068
[19] Yun, H., Kim, H., Jeong, Y.H. *et al.* Autoencoder-based anomaly detection of industrial robot arm using stethoscope based internal sound sensor. *J Intell Manuf* **34**, 1427–1444 (2023). https://doi.org/10.1007/s10845-021-01862-4
[20] Kullu, O., & Cinar, E. (2022). A deep-learning-based multi-modal sensor fusion approach for detection of equipment faults. *Machines*, *10*(11), 1105. https://doi.org/10.3390/machines10111105
[21] Marco, V. R., Kalkkuhl, J., Raisch, J., Scholte, W. J., Nijmeijer, H., & Seel, T. (2020). Multi-modal sensor fusion for highly accurate vehicle motion state estimation. *Control Engineering Practice*, *100*, 104409. https://doi.org/10.1016/j.conengprac.2020.104409

[22] Yuan, L., Andrews, J., Mu, H., Vakil, A., Ewing, R., Blasch, E., & Li, J. (2022). Interpretable passive multi-modal sensor fusion for human identification and activity recognition. *Sensors*, *22*(15), 5787. https://doi.org/10.3390/s22155787
[23] Petrich, J., Snow, Z., Corbin, D., & Reutzel, E. W. (2021). Multi-modal sensor fusion with machine learning for data-driven process monitoring for additive manufacturing. *Additive Manufacturing*, *48*, 102364. https://doi.org/10.1016/j.addma.2021.102364
[24] Lilan, L. I. U., Xiang, W. A. N., & Zenggui, G. A. O. (2023). An improved MPGA-ACO-BP algorithm and comprehensive evaluation system for intelligence workshop multi-modal data fusion. *Advanced Engineering Informatics*, *56*, 101980. https://doi.org/10.1016/j.aei.2023.101980
[25] Zhang, Y., Sheng, M., Liu, X. *et al.* A heterogeneous multi-modal medical data fusion framework supporting hybrid data exploration. *Health Inf Sci Syst* **10**, 22 (2022). https://doi.org/10.1007/s13755-022-00183-x
[26] Georgiou, T., Liu, Y., Chen, W. *et al.* A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision. *Int J Multimed Info Retr* **9**, 135–170 (2020). https://doi.org/10.1007/s13735-019-00183-w
[27] Shaukat, K., Luo, S., & Varadharajan, V. (2023). A novel deep learning-based approach for malware detection. *Engineering Applications of Artificial Intelligence*, *122*, 106030. https://doi.org/10.1016/j.engappai.2023.106030
[28] Wang, P., Fan, E., & Wang, P. (2021). Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern recognition letters*, *141*, 61-67. https://doi.org/10.1016/j.patrec.2020.07.042
[29] Priyasad, D., Fernando, T., Denman, S., Sridharan, S., & Fookes, C. (2021). Memory based fusion for multi-modal deep learning. *Information Fusion*, *67*, 136-146. https://doi.org/10.1016/j.inffus.2020.10.005
[30] Salama, E. S., El-Khoribi, R. A., Shoman, M. E., & Shalaby, M. A. W. (2021). A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition. *Egyptian Informatics Journal*, *22*(2), 167-176. https://doi.org/10.1016/j.eij.2020.07.005
[31] Tembhurne, J.V., Diwan, T. Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks. *Multimed Tools Appl* **80**, 6871–6910 (2021). https://doi.org/10.1007/s11042-020-10037-x
[32] Duan, J., Xiong, J., Li, Y., & Ding, W. (2024). Deep learning based multimodal biomedical data fusion: An overview and comparative review. *Information Fusion*, 102536. https://doi.org/10.1016/j.inffus.2024.102536
[33] Sun, H., Liu, J., Chai, S., Qiu, Z., Lin, L., Huang, X., & Chen, Y. (2021). Multi-modal adaptive fusion transformer network for the estimation of depression level. *Sensors*, *21*(14), 4764. https://doi.org/10.3390/s21144764
[34] Mousa, R., Taherinia, H., Abdiyeva, K., Bengari, A. A., & Vahediahmar, M. (2025). Integrating vision and location with transformers: A multimodal deep learning framework for medical wound analysis. *arXiv preprint arXiv:2504.10452*. https://doi.org/10.48550/arXiv.2504.10452
[35] Xie, W., Fang, Y., Yang, G., Yu, K., & Li, W. (2023). Transformer-based multi-modal data fusion method for COPD classification and physiological and biochemical indicators identification. *Biomolecules*, *13*(9), 1391. https://doi.org/10.3390/biom13091391
[36] Li, J., Yang, C., Ye, G., & Nguyen, Q. V. H. (2024). Graph neural networks with deep mutual learning for designing multi-modal recommendation systems. *Information Sciences*, *654*, 119815. https://doi.org/10.1016/j.ins.2023.119815
[37] Zhang, T., Chen, S., Wulamu, A. *et al.* TransG-net: transformer and graph neural network based multi-modal data fusion network for molecular properties prediction. *Appl Intell* **53**, 16077–16088 (2023). https://doi.org/10.1007/s10489-022-04351-0
[38] Zhou, T., Canu, S., & Ruan, S. (2020). Fusion based on attention mechanism and context constraint for multi-modal brain tumor segmentation. *Computerized Medical Imaging and Graphics*, *86*, 101811. https://doi.org/10.1016/j.compmedimag.2020.101811
[39] Liu, Y., Sun, H., Guan, W., Xia, Y., & Zhao, Z. (2022). Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework. *Speech Communication*, *139*, 1-9. https://doi.org/10.1016/j.specom.2022.02.006
[40] Zhou, Yang & Yang, Cai & Wang, Ping & Wang, Chao & Wang, Xinhong & Van, Nguyen Ngoc. (2024). ViT-FuseNet: MultiModal Fusion of Vision Transformer for Vehicle-Infrastructure Cooperative Perception. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3368404.
[41] Müller, S., Wengefeld, T., Trinh, T. Q., Aganian, D., Eisenbach, M., & Gross, H. M. (2020). A multi-modal person perception framework for socially interactive mobile service robots. *Sensors*, *20*(3), 722. https://doi.org/10.3390/s20030722

[42] Lin, K., Li, Y., Sun, J., Zhou, D., & Zhang, Q. (2020). Multi-sensor fusion for body sensor network in medical human–robot interaction scenario. *Information Fusion*, *57*, 15-26. https://doi.org/10.1016/j.inffus.2019.11.001
[43] Boroushaki, T., Dodds, L., Naeem, N., & Adib, F. (2023). FuseBot: mechanical search of rigid and deformable objects via multi-modal perception. *Autonomous Robots*, *47*(8), 1137-1154. https://doi.org/10.1007/s10514-023-10137-1
[44] Hou, R., Chen, G., Han, Y., Tang, Z., & Ru, Q. (2022). Multi-modal feature fusion for 3D object detection in the production workshop. *Applied Soft Computing*, *115*, 108245. https://doi.org/10.1016/j.asoc.2021.108245
[45] Xia, B., Zhou, J., Kong, F., You, Y., Yang, J., & Lin, L. (2024). Enhancing 3D object detection through multi-modal fusion for cooperative perception. *Alexandria Engineering Journal*, *104*, 46-55. https://doi.org/10.1016/j.aej.2024.06.025
[46] Lai, T. (2022). A review on visual-slam: Advancements from geometric modelling to learning-based semantic scene understanding using multi-modal sensor fusion. *Sensors*, *22*(19), 7265. https://doi.org/10.3390/s22197265
[47] Huang, Z., Sun, S., Zhao, J., & Mao, L. (2023). Multi-modal policy fusion for end-to-end autonomous driving. *Information Fusion*, *98*, 101834. https://doi.org/10.1016/j.inffus.2023.101834
[48] Xia, Y., Liu, S., Yu, Q., Deng, L., Zhang, Y., Su, H., & Zheng, K. (2023). Parameterized decision-making with multi-modal perception for autonomous driving. *arXiv preprint arXiv:2312.11935*. https://doi.org/10.48550/arXiv.2312.11935
[49] Wang, Y., Dong, B., Zhang, Y., Zhou, Y., Mei, H., Wei, Z., & Yang, X. (2023, October). Event-enhanced multi-modal spiking neural network for dynamic obstacle avoidance. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 3138-3148). https://doi.org/10.1145/3581783.3612147
[50] Du, C., Li, T., Liu, Y., Wen, Z., Hua, T., Wang, Y., & Zhao, H. (2021). Improving multi-modal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*. https://doi.org/10.48550/arXiv.2106.11059
[51] Ma, F., Li, Y., & Chen, M. (2024). Tactile texture recognition of multi-modal bionic finger based on multi-modal CBAM-CNN interpretable method. *Displays*, *83*, 102732. https://doi.org/10.1016/j.displa.2024.102732
[52] Wu, Y., Jiang, X., Fang, Z., Gao, Y., & Fujita, H. (2021). Multi-modal 3d object detection by 2d-guided precision anchor proposal and multi-layer fusion. *Applied Soft Computing*, *108*, 107405. https://doi.org/10.1016/j.asoc.2021.107405
[53] Wang, Y., Du, S., Xin, Q., He, Y., & Qian, W. (2024). Autonomous driving system driven by Artificial intelligence perception fusion. *Academic Journal of Science and Technology*, *9*(2), 193-198.
[54] de Oliveira, T. E. A., & da Fonseca, V. P. (2023). BioIn-Tacto: A compliant multi-modal tactile sensing module for robotic tasks. *HardwareX*, *16*, e00478. https://doi.org/10.1016/j.ohx.2023.e00478
[55] Tang, Z., Chen, G., Han, Y., Liao, X., Ru, Q., & Wu, Y. (2022). Bi-stage multi-modal 3D instance segmentation method for production workshop scene. *Engineering Applications of Artificial Intelligence*, *112*, 104858. https://doi.org/10.1016/j.engappai.2022.104858
[56] Weerakoon, K., Sathyamoorthy, A. J., Liang, J., Guan, T., Patel, U., & Manocha, D. (2022). Graspe: Graph based multimodal fusion for robot navigation in unstructured outdoor environments. *arXiv preprint arXiv:2209.05722*. https://doi.org/10.48550/arXiv.2209.05722
[57] Yin, C., Zhang, Q. A Multi-modal Framework for Robots to Learn Manipulation Tasks from Human Demonstrations. *J Intell Robot Syst* **107**, 56 (2023). https://doi.org/10.1007/s10846-023-01856-9