A Survey on Multimodal Data Fusion for Autonomous Collaborative Robots: Advances and real world challenges

Khalil ZARROUK
University of Sousse,
Ecole Nationale d'Ingénieurs de Sousse,
LATIS- Laboratory of
Advanced Technology and
Intelligent Systems, 4002,
Sousse, Tunisia
khalil.zarrouk@eniso.u-sousse.tn

Lamia RZOUGA HADDADA
University of Sousse,
Higher Institute of Applied
Science and Technology of Sousse,
LATIS- Laboratory of Advanced
Technology and Intelligent Systems, 4002,
Sousse, Tunisia
lamia.rzouga@issatso.u-sousse.tn

Sami GAZZAH
University of Sousse,
ISITCom, LATIS- Laboratory
of Advanced Technology and
Intelligent Systems, 4002, Sousse, Tunisia
sami.gazzah@gmail.com

***Abstract*—Collaborative robots (cobots) are revolutionizing industries, warfare, and smart cities. Multimodal data fusion (MMDF) enables a smart system to predict or decide based on the basis of multiple sensor input modalities. This survey comprehensively reviews MMDF techniques and their use cases for various cobots architectures and general approaches for multi-robot cooperation. By addressing future directions and challenges such as real-time processing and robustness in dynamic environments and identifying open research questions, this survey aims to guide future developments in the field, fostering innovation in robot collaboration.**

***Index Terms*—Multimodal data fusion, Collaborative robots, multiple sensors integration.**

I. Introduction

Multimodal Data Fusion (MMDF) is the process of combining data from different modalities to generate meaningful and comprehensible information [1]. cobots heavily rely on multiple data modalities from various sources, including cameras, IMUs, and LiDAR sensors. These robots are designed to operate alongside humans in diverse fields such as industry and services. By gathering data from multiple sensors, they perform critical decision-making tasks, including obstacle avoidance, emergency stops, and return-to-home functionalities similar to how DJI's Mavic UAVs (Unmanned Aerial Vehicles) respond when communication is lost [2]. cobots architectures can be classified into four main categories: centralized, decentralized, hybrid (semi-centralized), and distributed. The selected architecture directly influences the fusion method employed. The MMDF for cobots is a significant and rapidly growing research area within artificial intelligence (AI) and robotics. Figure 1 presents statistics on the increasing number of academic publications per 2 years from platforms such as IEEE Xplore, Springer, ACM Digital Library, ScienceDirect, and Wiley.

The MMDF for cobots presents as many future directions as challenges, including the lack of public datasets [3]. This paper focuses on MMDF techniques, cobots architectures and their existing fusion methods, public datasets used for collaborative SLAM (Simultaneous Localization And Mapping), and the challenges and future direction of MMDF for cobots, while trying to provide as much information as possible in a brief way.

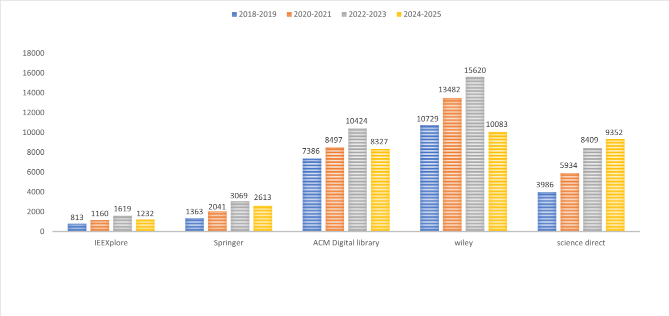

Fig. 1. The increase of paper per 2 years for each platform

This paper is structured as follows: Section 1 introduces the research domain. Section 2 discusses MMDF and cobots. Section 3 explores ML-based fusion techniques and the impact of DL on MMDF. Section 4 addresses fusion techniques for cobots. Section 5 outlines key challenges and future research directions. Finally, Section 6 concludes the study.

II. Multimodal data fusion and cobots

A. *Multimodal Data Fusion: Definition and related works*

The history of MMDF for cobots started in the 1980s, when the first simultaneous localization and mapping (SLAM) problem was proposed by Smith et al. in 1986 [4]. In the late 1990s, researchers began applying algorithms designed for single-robot systems t multi-robot systems, leading to significant advancements in this field. Around 2003, the introduction of cobots marked a turning point in the MMDF applications [5].

Hung et al. [6] designed a UGV (Unmanned Ground Vehicle) that fuses data from the LIDAR and RGB camera after performing a segmentation process on captured pictures.

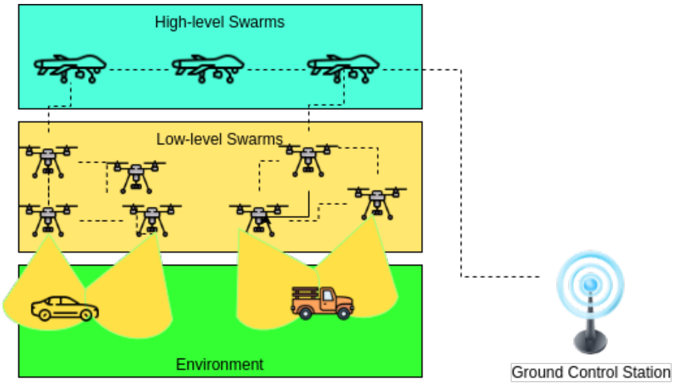

Fig. 2. Illustration of the ASIMUT project

The model fuses various types of data to make real-time predictions on decisions, including avoidance of obstacles. In cobots such as Industry 4.0, IBoT (Internet of Battlefield Things, which is a military application of IoT), data fusion is crucial and imposes additional fusion techniques based on the robot's distribution; Data can be heterogeneous or homogeneous, and sources can be another robot's input, prediction/decision.

In the ASIMUT (Aid to SItuation Management based on MUltimodal, MUltiUAVs, MUltilevel acquisition Techniques) project for surveillance services using heterogeneous, multilevel swarms of UAVs [7] which is described in Figure 2. The ASIMUT project is based on a hierarchical swarm system architecture that works as follows: High-Level Coordination Swarm (HLCS), UAVs with higher-level control and coordination roles to monitor the overall mission and provide commands to the low-level swarms, and act as intermediaries between the ground control station and low-level swarms. The Low-Level Swarms (LLS), UAVs that perform localized tasks, such as surveillance, target detection. They collaborate to perform their mission. The Ground Control Station (GCS) serves as the interface between human operators and the swarm system. The data fusion in that project is performed in three levels [8] as follows: Level 1, object assessment, consists of Identifying and tracking targets, also involves associating raw data to specific entities, like detecting and locating a vehicle. Level 2, situation assessment, which distinguishes the relationships between detected objects to assess the current situation.Level 3, impact assessment, which predicts the potential consequences or threats.

In [9], Ying Zhang et al. propose a framework for Air-Ground collaboration, where UAVs and UGVs coordinate to execute rescue missions focused on mapping and navigation. The system architecture is designed as follows: UAVs, equipped with sensors such as cameras, LiDAR, and thermal imagers, perceive and map the environment, while UGVs utilize these maps for path planning and navigation to carry out fire and rescue operations.

B. Architectures of cobots

Depending on their functionalities, cobots can be communicated in four major architectures (I) as the following:

- Centralized architectures: A single central controller manages all the robots relying on direct communication between central and related robots which can become a bottleneck. In [10], Barreto-Cubero et al. proposed a neural network-based sensor fusion system for service robots with a centralized architecture. Their hybrid deep convolution and recurrent neural network improved data processing. Ali et al. [11] presented flocking strategies for cobots emphasizing centralized control models for large-scale multi-robot systems.
- Decentralized architectures: The robots operate with local decision-making based on peer-to-peer communication. In [12], Tan et al. developed a decentralized navigation system for multi-robot exploration, using deep reinforcement learning to improve scalability and reduce communication overhead. Furthermore, Longa et al. [13] developed an autonomous multi-drone warehousing system using a decentralized multi-agent reinforcement learning approach showing robust performance across a fleet of three drones, highlighting the scalability of decentralized deep learning techniques in robotics.
- Hybrid architectures: It combines other architectures that can adapt between centralized and distributed modes. It is flexible for different tasks and situations. Tsang et al. [14], presented a warehouse multi-robot automation system where a centralized server handles task allocation while each robot computes its own local path planning. Liu et al. [15] proposed a hierarchical planning structure for a large-scale warehouse logistics system using centralized task allocation to assign tasks based on factors such as travel distance and current failures alongside local path planning
- Distributed architectures: A fully decentralized system with a greater emphasis on collaborative decision-making and shared information. Queralta et al. [16] proposed a framework where cobots use blockchain technology to coordinate and share information. Ceren et al. [17] developed a distributed architecture for agricultural robots that enables peer-to-peer coordination to reduce crops loss.

III. Fusion techniques

Current research is mainly categorized into traditional machine learning-based and deep learning-based approaches.

A. Traditional machine learning-based approaches

In the 1990s, with the appearance of machine learning, ML-based fusion models began to thrive, the goal was to derive insights from multimodal data to guide decision-making. It works by constructing multiple base classifiers (machine learning algorithms) and combining them to complete a learning task and solve a particular computational problem [18]. ML-based MMDF models can be categorized into three main

TABLE I
MAJOR COLLABORATIVE ROBOT ARCHITECTURES

Aspect	**Centralized**	**Decentralized**	**Hybrid**	**Distributed**
Control Location	Single central controller	Local controllers	Central + local	Fully distributed
Scalability	Limited	High	Moderate	High
Robustness	Low (single point of failure)	High	High	Very high
Latency	High	Low	Moderate	Moderate
Strengths	Global optimization and efficient task allocation	Robust to failures and scalable	Combines global perspective with local autonomy	Highly resilient and adaptive
Weaknesses	Prone to bottlenecks and single-point failures	Limited global awareness and complex coordination	Increased complexity	High communication overhead for consensus
Applications	Manufacturing, human-robot interaction	Swarm robotics, exploration	Warehousing, autonomous fleets	Agriculture, search-and-rescue missions
References	Barreto-Cubero et al (2021) [10], Ali et al. (2023) [11]	Tan et al. (2022) [12],Longa et al. (2024) [13]	Tsang et al. (2018) [14], Liu et al. (2019) [15]	Queralta et al. (2019) [16], Ceren et al. (2019) [17]

methods [19]: There are three types of fusion: early fusion (data-level fusion), intermediate fusion (feature-level fusion), and late fusion (decision-level fusion). These methods can be combined in a hybrid fusion method. Table II provides a comparison of these methods.

Most of the work was centered on feature engineering, relying on numerous hand-crafted extractors guided by prior knowledge—an approach that struggled to capture both the complementary and redundant relationships between modalities. This issue was solved with the appearance of deep learning. [20].

B. Deep Neural Network-based approaches

Since 2010, the use of the DL has led to outstanding results [21]. Deep Neural Network (DNN), the core of the DL-based methodology, showed superior performance by providing automated feature engineering [1]. The DNN architecture become more sophisticated to capture richer representations from multiple data sources. In these approaches, representation learning, modality fusion, and decision-making are often intertwined rather than applied sequentially to each modality. Consequently, fusion strategies have evolved beyond traditional early, intermediate, and late fusion toward more implicit, end-to-end methods. DNNs can learn complex relationships and patterns from multiple data inputs making them suitable for various tasks [22]. Table III describes some DNN-based methods.

IV. FUSION METHODS FOR COBOTS

A. MMDF architecture dependencies for cobots

For cobots, the fusion method depends heavily on the used architecture. The key considerations for this are:

- Sensor Configuration [5]: Using multiple heterogeneous sensors may require advanced fusion methods such as DL-based approaches or Bayesian networks. Also, real-time applications need fusion methods that offer low latency, such as Kalman filters and particle filters. In addition, the high variability in sensor input may require asynchronous fusion strategies.
- Real-Time Processing [24]: high-speed industrial tasks require lightweight and computationally efficient fusion methods, such as Kalman filter variants for linear fusion and decentralized fusion algorithms when computational power is distributed across multiple processors, cobots with high processing power can employ deep learning-based fusion methods.
- Task-Specific Requirements [25]: The architecture determines the cobot task, which influences the choice of fusion method, for example, for physical human-robot interaction, The use of MMDF methods to integrate visual, auditory, and tactile inputs for safe and seamless operation for direct interaction with humans.
- Communication and Data Flow (Centralized vs. Decentralized Fusion) [5]: Centralized architectures can use raw data fusion, while decentralized architectures rely on feature-level fusion.
- Scalability and Modularity [26]: Scalable architectures require adaptable fusion methods to additional sensors or data streams, like graph-based fusion for modular sensor networks, and transformer models for multisensor data integration.

B. Fusion techniques for cobots

Fusing inputs for cobots depends on various factors, including the chosen architecture and previously mentioned dependencies. In our comprehensive review of the literature, we explored numerous methods and approaches for this objective and summarized them in Table IV.

C. MMDF models Datasets for cobots

There are many public datasets designed for MMDF for cobots, and this paper focuses on C-SLAM (collaborative SLAM) datasets. Table V provides a comparison between widely recognized datasets. For time synchronization, HW refers to hardware solutions like GPS timing and hardware-based triggering mechanisms, while SW refers to software solutions like Network Time Protocol (NTP) or interpolation techniques.

V. CHALLENGES AND FUTURE DIRECTION

The integration of MMDF in cobots presents several challenges and future opportunities for researchers, which will be discussed in the following:

TABLE II
COMPREHENSIVE COMPARISON OF MULTIMODAL FUSION TECHNIQUES

Aspect	**Early Fusion**	**Intermediate Fusion**	**Late Fusion**
Definition	Combines raw data or features before learning	Fuses representations at hidden layers	Combines decisions from modality-specific models
Architecture	Single model processing concatenated features	Multiple streams with interaction layers	Separate models with fusion mechanism
Fusion Point	Input layer	Hidden layers	Output layer
Common Methods	Feature concatenation; Joint embeddings; Tensor fusion	Cross-attention; Multimodal transformers; Cross-modal gates	Weighted averaging; Majority voting; Learned aggregation
Advantages	Captures low-level interactions; Simple implementation; Joint optimization	Balanced feature interaction; Flexible architecture; Hierarchical learning	Modality independence; Easy to scale; Robust to missing data
Limitations	Curse of dimensionality; Modality synchronization; Scale differences	Complex architecture; Training difficulty; Computational cost	Misses early interactions; Independent optimization; Limited joint learning
Example Applications	Audio-visual speech recognition; Multimodal emotion recognition	Vision-language tasks; Complex multimedia analysis	Medical diagnosis; Multi-sensor systems
Related Works	Q. Zhang et al. (2023) [33]	Y. Li et al. (2023) [34]	R. Chen et al. (2023) [35]

TABLE III
COMPARISON OF DEEP NEURAL NETWORK METHODS FOR MULTIMODAL DATA FUSION

Category	**Description**	**Applications**	**Related Work**
Encoder-Decoder Methods	Models use encoder networks to extract high-level features and decoder networks to generate predictions	Image segmentation, language translation	Couprie et al., 2013 [36]
Attention Mechanisms	Assign weights to input data components to focus on the most relevant information	Visual Question Answering, video captioning	Vaswani et al., 2017 [37]
Graph Neural Networks	Leverage graph structures to capture relationships between data points	RGB-depth scene classification	Lotfi et al., 2024 [38]
Generative Networks	Generate or reconstruct data by modeling underlying distributions	Missing data imputation	Gao et al., 2024 [39]
Constraint-Based Methods	Learn separate representations while enforcing semantic alignment	Multimodal sentiment analysis	Zadeh et al., 2017 [40]
Visual Geometry Group Neural Network (VGG19)	Fuse imagery data implicitly through pooling layers	Style Features Extraction	Haddada et al. 2024 [8]

TABLE IV
OVERVIEW OF SENSOR FUSION TECHNIQUES IN COLLABORATIVE ROBOTICS

Fusion approach	**Technique**	**Description**	**Key Applications**
Probabilistic Methods	Kalman Filtering [27]	Combines sensor data to estimate robot states.	State estimation, dynamic tracking
Probabilistic Methods	Bayesian Fusion [5]	Uses Bayesian probability to combine noisy sensor inputs and enable adaptive behavior	Human-robot collaboration, mapping
Probabilistic Methods	Covariance Intersection (CI) [28]	Fuses data with unknown correlations using weighted covariance matrices.	Decentralized SLAM, multi-robot systems
Machine Learning	KalmanNet [29]	Hybrid Kalman filter + neural networks for non-Gaussian noise adaptation.	Dynamic environments, navigation
Machine Learning	Reinforcement learning [30]	Learns optimal fusion policies through trial-and error feedback.	Adaptive manipulation, obstacle avoidance
Decentralized Fusion	Heterogeneous State CF (HS-CF) [31]	Tracks dependencies via factor graphs and covariance deflation for conservativeness.	Multi-robot tracking, cooperative localization
Decentralized Fusion	Factor Graph-based DDF [32]	Uses factor graphs for scalable decentralized inference in dynamic systems.	Distributed mapping, target tracking

A. Challenges

- **Data heterogeneity:** Managing variations in data resolution, temporal alignment, and modality-specific noise; This can be solved by developing frameworks for efficient normalization, synchronization, and preprocessing of multimodal data [49].
- **Sensor calibration and drift:** Long-term operation leads to sensor drift, which degrades the fusion accuracy, requiring the implementation of self-calibration algorithms for consistent sensor performance [50].
- **Fusion complexity and real-time performance:** Balancing the computational cost of complex MMDF models with the real-time requirements of cobots requires lightweight models and approximation techniques for faster inference [50].
- **Safety and robustness:** Ensuring that cobots can operate safely even when sensor data is incomplete or corrupted. This can be solved by developing robust multimodal models that can handle uncertainty and failures [50].
- **Energy efficiency:** Since there is high computational re-

TABLE V
COMPARATIVE TABLE OF DATASETS FOR COLLABORATIVE SLAM (C-SLAM)

Dataset	**Platform**	**Sensors**	**Time Sync.**	**Trajectory Overlap**	**Environment**	**Ground Truth**
KITTI [41]	Car	Camera, IMU, LiDAR, GPS	HW	Large	Outdoor	GNSS/INS
EUROC MAV [42]	UAV	Camera, IMU	SW	Restricted	Indoor	Motion Capture
Oxford Robotics Car [43]	Car	Camera, IMU, LiDAR, GPS	HW	Large	Outdoor	GNSS/INS
TUM RGB-D [44]	Handheld	Camera, RGB-D	HW	Restricted	Indoor	3D Scanner
MulRan [45]	Car	Radar, LiDAR, GPS	SW	Large	Outdoor	GNSS/INS
NCLT [46]	Segway Robot	Camera, IMU, LiDAR, GPS	HW	Large	Outdoor	GNSS/INS
KAIST Urban [47]	Car	Camera, IMU, LiDAR	HW	Restricted	Outdoor	GNSS/INS
S3E [48]	UGVs, UAVs, Handheld	Camera, IMU, LiDAR, UWB	HW	Restricted	Indoor, Outdoor	3D Scanner

quirements for battery-operated multimodal fusion strain cobots, therefore, models must be optimized for low power consumption, leveraging edge AI for resource-constrained environments [51].

B. Future Directions

- **Development of real-time multimodal fusion algorithms:** Enabling cobots to process and fuse multimodal data in real-time by relying on advanced neural architectures like transformer-based models or graph neural networks (GNNs) that process multi-source data efficiently, which will improve real-time response and then cobots interactions in dynamic environments [52].
- **Context-aware fusion:** Introducing context-aware mechanisms to prioritize specific modalities based on the task or environment by using adaptive weighting schemes and attention mechanisms that dynamically adjust data importance, which increases cobots reliability and adaptability [53].
- **Scalable multimodal fusion for multi-agent systems:** Developing MMDF techniques to support multi-robot collaboration, ensuring seamless integration of data from multiple cobots and sensors, which can be applied on robots swarm or distributed SLAM, enabling large-scale collaboration for tasks like search-and-rescue operations [54].
- **Cloud-edge integration for collaborative fusion:** Relying on edge computing and cloud services to enable distributed MMDF in collaborative scenarios, which can be applied in resource-constrained environments where the computational load needs to be shared, that expands the scalability and computational capability of MMDF systems [54].
- **Action-reflex mechanism:** Developing an action-relfex mechanism allows cobots to deal with emergencies, such as unobserved obstacles, moving vehicles, and extreme weather, similar to muscle-conditioned reflex which organizes local muscles to avoid hazards in the first response without delaying passage through the brain [55].

VI. CONCLUSION

This survey has examined the current state, challenges, and future directions of MMDF in cobots. Through our analysis, several key findings emerge. First, the evolution of MMDF techniques from traditional machine learning approaches to deep learning has enhanced cobots enabling more robust and adaptive fusion capabilities. Second, the choice of fusion architecture affects the performance and capabilities of the system. Third, the integration of multiple sensors and data streams presents both opportunities and challenges. The state of the art presented in this survey helps researchers and practitioners working on next-generation cobots systems. As the field continues to advance, the integration of new sensing modalities, improved fusion algorithms, and more sophisticated architectural approaches will drive innovation in collaborative robotics.

REFERENCES

[1] F. Zhao, C. Zhang, and B. Geng, “Deep multimodal data fusion,” *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–36, 2024.

[2] DJI, ”Understanding Mavic UAV Functions,” DJI Support, 2025.

[3] C. Zhang, Z. Yu, X. Wang, and C. Deng, “End-to-end lower limb motion recognition with modal-channel attention and dual-branch adaptive fusion network via sEMG and kinematic data,” in *Proc. 7th Int. Conf. Electron. Technol. (ICET)*, 2024, pp. 1053–1058.

[4] R. C. Smith and P. Cheeseman, “On the representation and estimation of spatial uncertainty,” *The International Journal of Robotics Research*, vol. 5, no. 4, pp. 56–68, 1986.

[5] W. Chen, X. Wang, S. Gao, G. Shang, C. Zhou, Z. Li, C. Xu, and K. Hu, “Overview of multi-robot collaborative SLAM from the perspective of data fusion,” *Machines*, vol. 11, no. 6, p. 653, 2023.

[6] X. Huang, H. Deng, W. Zhang, R. Song, and Y. Li, “Towards multi-modal perception-based navigation: A deep reinforcement learning method,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4986–4993, 2021.

[7] P. Bouvry *et al.*, “Using heterogeneous multilevel swarms of UAVs and high-level data fusion to support situation management in surveillance scenarios,” in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*, 2016, pp. 424–429.

[8] L. R. Haddada, F. M. Rmida, W. Ouarda, I. K. Kallel, R. Maalej, S. Masmoudi, A. M. Alimi, and N. E. B. Amara, “A benchmark tetra-modal biometric score database,” *Biomedical Signal Processing and Control*, vol. 98, p. 106778, 2024.

[9] Y. Zhang *et al.*, “Air-ground collaborative robots for fire and rescue missions: Towards mapping and navigation perspective,” *arXiv preprint arXiv:2412.20699*, 2024.

[10] A. J. Barreto-Cubero, A. Gómez-Espinosa, J. A. Escobedo Cabello, E. Cuan-Urquizo, and S. R. Cruz-Ramírez, “Sensor data fusion for a mobile robot using neural networks,” *Sensors*, vol. 22, no. 1, Art. no. 305, 2022.

[11] Z. A. Ali, E. H. Alkhammash, and R. Hasan, “State-of-the-art flocking strategies for the collective motion of multi-robots,” *Machines*, vol. 12, no. 10, p. 739, 2024.

[12] A. H. Tan, F. P. Bejarano, Y. Zhu, R. Ren, and G. Nejat, “Deep reinforcement learning for decentralized multi-robot exploration with macro actions,” *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 272–279, 2022.

[13] M. E. Longa, Z. Wei, A. Tsourdos, and G. Inalhan, “Autonomous multi-drone warehousing through deep reinforcement learning and predictive potential fields,” *Preprint*, 2024.

[14] K. F. E. Tsang, Y. Ni, C. F. R. Wong, and L. Shi, “A novel warehouse multi-robot automation system with semi-complete and computationally efficient path planning and adaptive genetic task allocation algorithms,” in *Proc. 15th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, 2018, pp. 1671–1676.

[15] Z. Liu, H. Wang, S. Zhou, Y. Shen, and Y.-H. Liu, “Coordinating large-scale robot networks with motion and communication uncertainties for logistics applications,” *arXiv preprint arXiv:1904.01303*, 2019.

[16] J. P. Queralta and T. Westerlund, “Blockchain-powered collaboration in heterogeneous swarms of robots,” *arXiv preprint arXiv:1912.01711*, 2019.

[17] R. Ceren, S. Quinn, and G. Raines, “Towards a decentralized, autonomous multiagent framework for mitigating crop loss,” *arXiv preprint arXiv:1901.02035*, 2019.

[18] F. Jiang, X. Yu, J. Du, D. Gong, Y. Zhang, and Y. Peng, “Ensemble learning based on approximate reducts and bootstrap sampling,” *Information Sciences*, vol. 547, pp. 797–813, 2021.

[19] L. R. Haddada, B. Dorizzi, and N. Essoukri Ben Amara, ”Watermarking signal fusion in multimodal biometrics,” in *Proc. Int. Image Processing, Applications and Systems Conf. (IPAS)*, 2014, pp. 1–6.

[20] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[21] Y. Dorai, F. Chausse, S. Gazzah, and N. E. B. Amara, “Multi target tracking by linking tracklets with a convolutional neural network,” in *Proc. 12th Int. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, vol. 6: VISAPP, 2017, pp. 492–498.

[22] S. R. Keskin, A. Gençdoğmuş, B. Yıldırım, G. Doğan, and Y. Öztürk, “DNN and CNN approach for human activity recognition,” in *2020 7th International Conference on Electrical and Electronics Engineering (ICEEE)*, pp. 254–258, 2020.

[23] K. Simonyan, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
IEEE Access, vol. 9, pp. 26208–26220, 2020. IEEE.

[24] M. Mukherjee, A. Banerjee, A. Papadimitriou, S. S. Mansouri, and G. Nikolakopoulos, “A decentralized sensor fusion scheme for multi sensorial fault resilient pose estimation,” *Sensors*, vol. 21, no. 24, p. 8259, 2021.

[25] M. Costanzo, G. De Maria, G. Lettera, and C. Natale, “A multimodal approach to human safety in collaborative robotic workcells,” *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 2, pp. 1202–1216, 2021. IEEE.

[26] D. Sani and S. Anand, “Graph-Based Multi-Modal Sensor Fusion for Autonomous Driving,” *arXiv preprint arXiv:2411.03702*, 2024.

[27] O. Dagan and N. R. Ahmed, “Conservative filtering for heterogeneous decentralized data fusion in dynamic robotic systems,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2022, pp. 5840–5847.

[28] S. J. Julier and J. K. Uhlmann, “A non-divergent estimation algorithm in the presence of unknown correlations,” in *Proc. Amer. Control Conf.*, vol. 4, 1997, pp. 2369–2373.

[29] G. Revach, N. Shlezinger, X. Ni, A. L. Escoriza, R. J. G. Van Sloun, and Y. C. Eldar, “KalmanNet: Neural network aided Kalman filtering for partially known dynamics,” *IEEE Trans. Signal Process.*, vol. 70, pp. 1532–1547, 2022.

[30] W. Chen, S. Zhou, Z. Pan, H. Zheng, and Y. Liu, “Mapless collaborative navigation for a multi-robot system based on the deep reinforcement learning,” *Appl. Sci.*, vol. 9, no. 20, p. 4198, 2019.

[31] H. Li and F. Nashashibi, “Cooperative multi-vehicle localization using split covariance intersection filter,” *IEEE Intell. Transp. Syst. Mag.*, vol. 5, no. 2, pp. 33–44, 2013.

[32] A. Cunningham, M. Paluri, and F. Dellaert, “DDF-SAM: Fully distributed SLAM using constrained factor graphs,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2010, pp. 3025–3030.

[33] Q. Zhang, X. Li, and J. Wang, ”MMFusion: Multimodal fusion with adaptive attention for emotion recognition,” IEEE Trans. Multimedia, vol. 25, pp. 2145-2159, Apr. 2023.

[34] Y. Li, S. Zhang, and R. Wang, ”UniperceiveR: Universal perception using transformers,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023, pp. 1205-1214.

[35] R. Chen, Y. Liu, and Z. Wang, ”AdaFusion: Adaptive multimodal fusion for robust HAR,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 6, pp. 7234-7248, Jun. 2023.

[36] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, “Indoor semantic segmentation using depth information,” *arXiv preprint arXiv:1301.3572*, 2013.

[37] A. Vaswani, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, 2017.

[38] S. Lotfi, M. Mirzarezaee, M. Hosseinzadeh, and V. Seydi, “Detection of rumor conversations in Twitter using graph convolutional networks,” *Applied Intelligence*, vol. 51, pp. 4774–4787, 2021.

[39] X. Gao, F. Shi, D. Shen, and M. Liu, “Task-induced pyramid and attention GAN for multimodal brain image imputation and classification in Alzheimer’s disease,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 36–43, 2021.

[40] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” *arXiv preprint arXiv:1707.07250*, 2017.

[41] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 3354–3361.

[42] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, “The EuRoC micro aerial vehicle datasets,” *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, 2016.

[43] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 year, 1000 km: The Oxford RobotCar dataset,” *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, 2017.

[44] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 year, 1000 km: The Oxford RobotCar dataset,” *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[45] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, “Mulran: Multimodal range dataset for urban place recognition,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2020, pp. 6246–6253.

[46] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, “University of Michigan North Campus long-term vision and lidar dataset,” *Int. J. Robot. Res.*, vol. 35, no. 9, pp. 1023–1035, 2016.

[47] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, “Complex urban dataset with multi-level sensors from highly diverse urban environments,” *Int. J. Robot. Res.*, vol. 38, no. 6, pp. 642–657, 2019.

[48] L. Chen, G. Li, W. Xie, J. Tan, Y. Li, J. Pu, L. Chen, D. Gan, and W. Shi, “A Survey of Computer Vision Detection, Visual SLAM Algorithms, and Their Applications in Energy-Efficient Autonomous Systems,” *Energies*, vol. 17, no. 20, 2024.

[49] M. Forlini, M. Babcinschi, G. Palmieri, and P. Neto, “D-RMGPT: Robot-assisted collaborative tasks driven by large multimodal models,” *arXiv preprint arXiv:2408.11761*, 2024.

[50] H. Liu, T. Fang, T. Zhou, and L. Wang, “Towards robust human-robot collaborative manufacturing: Multimodal fusion,” *IEEE Access*, vol. 6, pp. 74762–74771, 2018.

[51] H.-A. Rashid and T. Mohsenin, “TinyM^2 Net-V3: Memory-Aware Compressed Multimodal Deep Neural Networks for Sustainable Edge Deployment,” *arXiv preprint arXiv:2405.12353*, 2024.

[52] P. Fan and Q. Wu, “Advances in computer AI-assisted multimodal data fusion techniques,” *Applied Mathematics and Nonlinear Sciences*, vol. 9, Nov. 2024.

[53] Y. Zhang, A. V. Malawade, X. Zhang, Y. Li, D. Seong, M. A. Al Faruque, and S. Huang, “CARMA: Context-Aware Runtime Reconfiguration for Energy-Efficient Sensor Fusion,” in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, 2023, pp. 1–6.

[54] Z. Cai, X. Du, T. Huang, T. Lv, Z. Cai, and G. Gong, “Robotic Edge Intelligence for Energy-Efficient Human–Robot Collaboration,” *Sustainability*, vol. 16, no. 22, 2024.

[55] Z. Zhao, B. He, W. Luo, and R. Liu, “Collective conditioned reflex: A bio-inspired fast emergency reaction mechanism for designing safe multi-robot systems,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10985–10990, 2022.