Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.3171

2025, 44(1) 75-81

Deep learning for multi-modal data fusion in IoT applications

Anila Saghir [a, *], Anam Akbar [b], Asma Zafar [c], Asif Hasan [a]

[a] *Department of Telecommunication Engineering, Sir Syed University of Engineering & Technology, Karachi*

[b] *Department of Computer Science, Sir Syed University of Engineering & Technology, Karachi*

[c] *Department of Mathematics, Sir Syed University of Engineering & Technology, Karachi*

[*] Corresponding author: Anila Saghir, Email: asaghir@ssuet.edu.pk

Received: 18 January 2024, Accepted: 26 December 2024, Published: 01 January 2025

KEYWORDS

Multimodal Fusion

Semantic Segmentation

Internet of Things

Deep Convolution Networks

Generative Modals

ABSTRACT

With the rapid changes in technology, the Internet of Things (IoT) has also emerged with many diverse applications. A massive amount of data is generated and processed through the IoT-based sensors from these applications every day. This sensor-based data is categorized as either structured or unstructured data. Structured data is simpler to process, while the processing of unstructured data is complex, due to its diverse modalities. In IoT applications such as autonomous navigation, environmental monitoring and smart surveillance, semantic segmentation is required, and it relies on detailed scene understanding. The single-modal data like RGB, thermal or depth images fails to provide this detailed information independently. This research proposes a robust solution by fusing the multimodal data and employing a deep learning-based hybrid architecture that incorporates a generative model with a deep convolutional network. The unified model fuses RGB, thermal and depth images for semantic segmentation to improve the accuracy and reliability. The successful results validate the effectiveness of the proposed technique.

1. Introduction

The origin of the Internet of Things (IoT) was to connect electronic devices for uninterrupted communication and data gathering. The evolution of smart IoT gained momentum after 2014, driven by the development of wireless sensor networks, reshaping various industries, and prompting the adoption of IoT platforms for economic expansion [1].

The potential increase in the IoT-based devices that lead to IoT applications is anticipated to reach around 75 billion by 2025 and could potentially escalate to approximately 140 billion by 2030 if the observed growth rate persists [1-2]. The IoT sector is supported by the integration of emerging technologies like Edge, Fog, and Cloud Computing. These technologies enhance data processing, decision-making, and analytics, signifying the importance of data fusion in many IoT-based applications [3].

The IoT applications generate a massive amount of data from various sources. To use the data for the decision-making process or analytics, the integration or fusion of the data is required. Thus, the integration of the data for extracting useful information from different sources becomes a significant hurdle in IoT systems, requiring effective approaches to fuse, process, and extract valuable information successfully [4].

Various IoT applications have a distributed and heterogeneous environment, particularly with diverse systems and data represented in various feature spaces, posing a substantial challenge in data fusion [5-7].

Specifically, in semantic segmentation-based applications, unimodal data like RGB, thermal, or depth alone, often fails to provide detailed scene understanding. RGB data usually struggles in poor lighting conditions, while thermal and depth

modalities lack color and texture information independently. These limitations in working with unimodal data decrease the segmentation accuracy of the overall system and highlight the need to use fused multimodalities to improve IoT applications that require robust scene interpretations [8-10].

Challenging issues such as heterogeneity in data frames, data size expansion, and network unreliability make analyzing IoT data difficult and impact the efficiency and accuracy of various applications [11-12]. Various mathematical theories and probability-based approaches exist for data fusion but with very complex models [13-20]. With the advancement of technology, multimedia data experiences explosive growth, spanning text, image, audio, and video modalities. The conventional model for processing that type of data struggles to accommodate the progressively diverse nature of multimedia data [21-22]. These challenges are particularly becoming more critical in IoT applications like autonomous vehicles, smart surveillance and environmental monitoring, where accurate and reliable segmentation is highly required [23-24].

This massive amount of raw data requires intelligent management, forming the foundation for multimodal learning in a novel data fusion paradigm. The data can be combined from multiple sources and forms into a unified form, known as multimodal data fusion, facilitating simplified data representation for further processing. Effective multimodal data fusion is dependent upon the choice of an appropriate fusion technique [25]. General fusion techniques are broadly classified as: data-level fusion, involving synchronization, buffering, de-noising, and normalization; feature-level fusion (Early fusion), including feature normalization and selection; and decision-level fusion, applied after classifiers [26-28].

The challenges posed by multimodal data, characterized by high volume, variety, and veracity, highlight the need for advanced data fusion methods. Using deep diverse neural architectures in multimodal fusion techniques can lead to improved outcomes [29-30]. Deep neural architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and generative models such as auto-encoders and Generative Adversarial Networks (GAN), using a hierarchical computational model, capture multilevel abstract representations of data for improved fusion [31].

Research in the medical domain, with late fusion-based deep learning models to fuse multimodal data, has been carried out for the analysis of various disease patterns [32-34]. These studies explain that combining fusion models like late fusion with advanced neural networks like deep neural networks (DNNs) and convolutional neural networks (CNNs) significantly increases the accuracy of medical images and electronic health records.

Fusion methods, especially the data and feature levels, are gaining popularity for industrial diagnosis, and show a preference for Deep Neural Networks (DNNs) with the concatenation of both data and features fusion. However, it is challenging to manage the computational efforts and maintain the data quality during data processing [35]. 1D CNN models are also being used for the integration of multimodal and multiresolution signals in image applications. Yet, their limitation to 1D signals prompts the necessity for a denser model [36].

In IoT-based segmentation applications, the integration of diverse sensing modalities like 3D Light Detection and Ranging (LiDARs), RGB-Depth (RGB-D), and thermal cameras significantly enhances scene understanding by reducing uncertainty, especially in complex scenarios. By combining the data from these different visual, LiDAR, and thermal, provides richer spatial and contextual information and significantly improves the performance of learning models compared to using unimodal approaches [37-39].

This research presents an integrated fused encoder-decoder-based fully CNN architecture to facilitate the fusion of multi-modal data. The research primarily uses the power of Convolutional Neural Networks (CNNs) and generative models (Variational Auto-Encoder (VAEs)) to develop an integrated fused feature, extracting complex patterns from three distinct sources: colour, thermal, and depth sensor-based images. The goal is to achieve an effective multimodal fusion technique to address the limitations of unimodal data in semantic segmentation for IoT applications.

The remainder of the paper is organized as follows: A description of the proposed model is explained in Section 2, Sections 3 and 4 discuss the results and conclusions, and future work is presented in Section 5.

2. Unified Fusion Model

To resolve the multimodal data integration problem in IoT application of semantic segmentation through RGB, thermal and depth sensor-based images have been considered.

The proposed model follows a generative approach, integrating an auto-encoder with deep CNNs. The model is influenced by the VGG-16 architecture, with modifications in the last two fully connected layers. The auto-encoder is used to

compress the multimodal data into a latent space before decoding it for final segmentation.

The proposed scheme is divided into two main sections, both sections apply convolutional neural networks by integrating encoding and decoding approaches.

2.1 CNN Encoder

The encoder part of the proposed model uses the concept of concurrency on multiple data modalities. Each modality is encoded as a separate network. The encoder concurrently extracts features from different modalities and fuses them into a unified framework.

The fusion is applied at the feature level to integrate the multimodal data into the unified framework. Fusion occurs in two stages. In the first stage, the model performs a summation operation on the feature maps of the RGB and thermal images to aggregate their features. In the second stage, the summated result is concatenated with the depth values of the image.

The overall fused feature map is then passed through the max pooling layer followed by a convolutional layer to further extract the features. As the neural network progresses into deeper layers, the output passes through more convolutional and pooling layers.

The model applies batch normalization after each convolutional layer to minimize covariance. The normalized output is then passed to the ReLU activation function. The proposed encoder is shown in the Fig. 1.

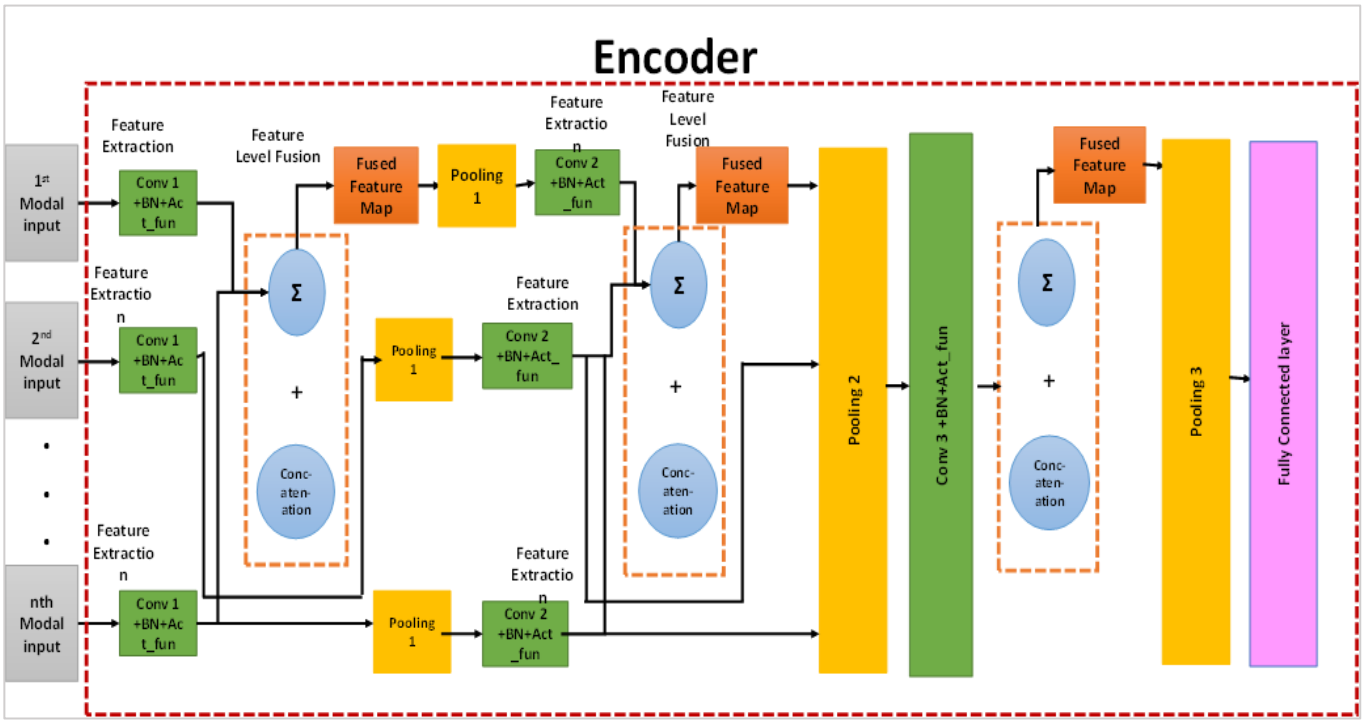

Fig. 1. Unified Fused CNN-based Encoder

2.2 CNN Decoder

The decoder part of the proposed technique performs the un-pooling (Max), deconvolution, and batch normalization and applies the SoftMax activation function for the final segmentation. The proposed decoder is shown in Fig. 2.

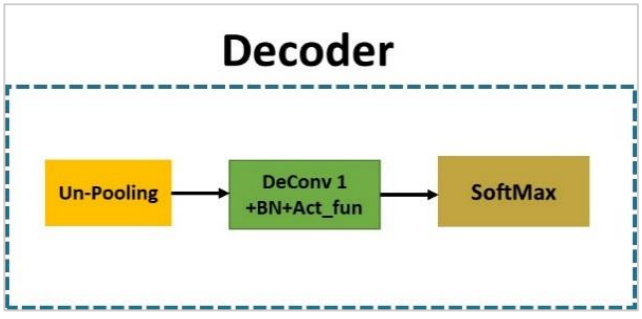

Fig. 2. Unified Fused CNN Based Decoder

2.3 Unified Fusion Algorithm

The detailed methodology is defined as:

- Input Parameters: VAP Trimodal People Segmentation Dataset from Kaggle has been taken. The collected data have 3 data modalities: RGB images (680x480), Thermal images (680x480), and Depth sensor data (1000-3300).
- Pre-processing: Normalized (depth images) and synchronized collected data (Registration File) to create a standardized dataset for effective fusion.
- Fused Deep Convolutional Encoder-Decoder: Integrated CNN and Autoencoder to get Fused Deep Convolutional Encoder-Decoder

Implemented concurrent encoding of data modalities.

- Feature Level Fusion: Applied feature-level fusion using both summation and concatenation based on modalities.
- Model Training: Trained and optimized the model using the pre-processed dataset and validated the model's performance through testing.
- Evaluation: Evaluated the model using performance metrics (accuracy).

  Accuracy = Number of Correct Predictions/Total number of predictions.

3. Results

An open access, multimodal dataset, named "VAP Trimodal People Segmentation" has been used for the training, testing and validation of the proposed technique. This dataset contains data from three different peoples of different activities captured from multimodal cameras, RGB, Thermal and depth sensors. The data set contains 5724 annotated frames for three scenes in the indoor setting.

Well-known performance metrics precision (P), recall (R), general Accuracy (GA) and class average accuracy (CAA) have been used to evaluate the overall performance of the trained model and its results across 09 classes. The proposed network is evaluated through extensive simulations. The training process was conducted for 50 epochs and was ended based on the validation loss.

Precision (P) is calculated as:

$$P=\frac{True\ Pred}{True\ Pred+False\ Pred}$$

Where precision (P) provides the estimation of the proportion of correct predictions among all positive predictions.

$$GA=\frac{1}{N}\sum_{i=1}^{N}P$$

Where, N=number of classes, and general accuracy (GA) represents the overall mean precision across all classes.

Recall (R) is calculated as:

$$R=\frac{True\ Pred}{True\ Pred+False\ Neg}$$

Where recall (R) is calculated to evaluate the proportion of correctly identified values vs actual values in the dataset.

$$CAA=\frac{1}{K}\sum_{i=1}^{K}R$$

Where, N=number of classes, and class average accuracy (CAA) represent the mean recall across all classes.

The proposed scheme is compared with Seg-Net Basic [39] and used as a baseline comparison. Also, segmentation using only RGB images (unimodal) has been implemented to compare and validate the results of the multimodal approach.

Table 1 shows the class-wise comparison of the quantitative results on the VAP Trimodal dataset for all three approaches, Seg-Net basic, Unimodal and proposed Unified Multimodal scheme, using precision, recall, general accuracy, and class average accuracy. The results demonstrate the improvement achieved by integrating the multimodal data for semantic segmentation.

Table 1

Comparison of quantitative results on the VAP Trimodal dataset

Class	Model	Precision (P)%	Recall (R)%	GA %	CAA %
Table	Unimodal (RGB only)	78.2	73.5	-	-
	Seg-Net Basic	90.1	83.2		-
	Proposed Unified Fusion	93.2	76.3	-	-
Laptop	Unimodal (RGB only)	72.4	69.0	-	-
	Seg-Net Basic	83.2	68.5	-	-
	Proposed Unified Fusion	91.4	79.0	-	-
Long Chair	Unimodal (RGB only)	68.5	62.0	-	-
	Seg-Net Basic	40.2	44.7	-	-
	Proposed Unified Fusion	84.2	81.3	-	-
Chair	Unimodal (RGB only)	36.0	31.4	-	-
	Seg-Net Basic	92.2	85.3	-	-
	Proposed Unified Fusion	95.3	92.1	-	-
Stand	Unimodal (RGB only)	45.3	46.1	-	-
	Seg-Net Basic	55.0	50.2	-	-

	Proposed Unified Fusion	75.2	72.8	-	-
Pot	Unimodal (RGB only)	58.4	54.5	-	-
	Seg-Net Basic	75.0	71.3	-	-
	Proposed Unified Fusion	80.3	77.1	-	-
Window	Unimodal (RGB only)	71.3	63.8	-	-
	Seg-Net Basic	82.7	79.3	-	-
	Proposed Unified Fusion	93.6	90.5	-	-
Ceiling	Unimodal (RGB only)	50.2	45.0	-	-
	Seg-Net Basic	55.0	50.1	-	-
	Proposed Unified Fusion	78.5	75.4	-	-
	Unimodal (RGB only)	45.0	39.5	-	-
Light	Seg-Net Basic	55.0	50.1	-	-
	Proposed Unified Fusion	78.5	75.4	-	-
	Unimodal (RGB only)	-	-	58. 9	53.9
	Seg-Net Basic	-	-	60.2	79.6
	Proposed Unified Fusion	-	-	80.6	85.3

Fig. 3a represents the accuracy of the model throughout the training and testing phases, which clearly shows the model's performance over time and epochs. Fig. 3b represents the loss calculations for the training and validation processes.

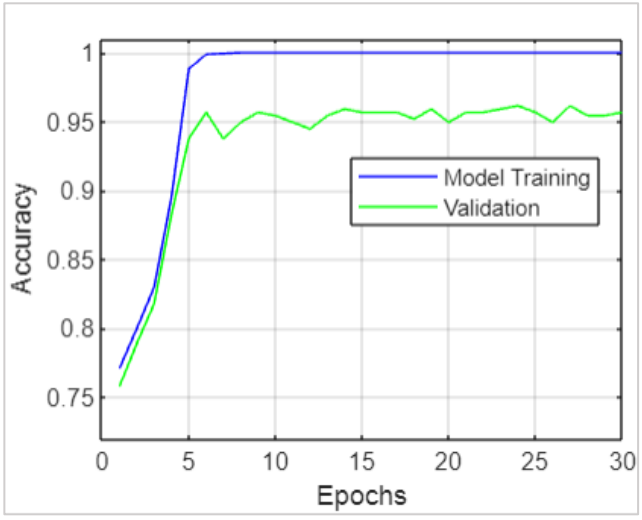

Fig. 3a. Accuracy Of Model During Training And Testing Phase

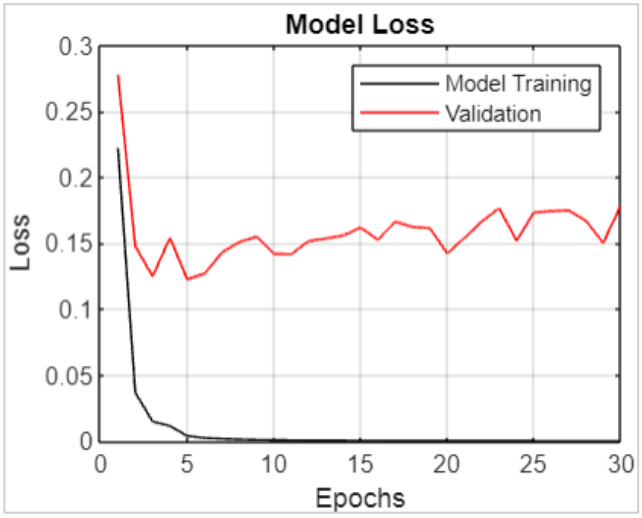

Fig. 3b. Loss During Training And Testing Phase

4. Conclusion

The paper presents a solution for fusing multimodal data from RGB, Thermal, and Depth sensors to improve semantic segmentation accuracy. The proposed fusion technique worked at the feature level, to fuse the feature maps of three modalities, which will effectively increase the overall accuracy of semantic segmentation.

The unified fused model employs a deep Convolutional Neural Network (CNN) architecture, incorporating a generative model (Autoencoder-Decoder). The feature maps from RGB and Thermal modalities are initially fused using a summation, which then then combined with depth information to further refine the segmentation process.

The unified fused model has efficiently performed semantic segmentation and offered remarkable results in comparison with existing fusion-based approaches.

5. Future Work

The positive outcomes achieved through the simulation validate the approach and methodology. However, further refinement is necessary to ensure the model's adaptability and reliability across diverse environmental conditions. The proposed model can be enhanced for numerous real-world scenarios like varying lighting, dynamic object movements and discrepancies in sensors.

6. Acknowledgment

The authors would like to acknowledge the Sir Syed University of Engineering and Technology for the overwhelming support and assistance throughout this work.

7. Reference

[1] M. S. Kaiser, et al., Information and Communication Technology for Competitive Strategies (ICTCS 2020), Springer Nature Singapore, pp. 393-402, 2020. doi: 10.1007/978-981-16-0882-7.

[2] W. T. Toor, M. Alvi, and M. Agiwal, "Combined access barring scheme for IoT devices using Bayesian estimation", Electronics, vol. 9, no. 12, pp. 2191, 2020, doi: 10.3390/electronics9122191.

[3] R. Krishnamurthi, et al., "An overview of IoT sensor data processing, fusion, and analysis techniques", Sensors, vol. 20, no. 21, pp. 6076, 2020.

[4] M. Abu-Elkheir, et al., "Data management for the Internet of Things: Design primitives and solution", Sensors, vol. 13, no. 11, pp. 15582-155612, Nov. 2013.

[5] J. Santos, T. Wauters, B. Volckaert, and F. De Turck, "Resource provisioning for IoT application services in smart cities", Proc. 13th Int. Conf. Network and Service Management (CNSM), 2017, pp. 1–9.

[6] S. Boulkaboul and D. Djenouri, "DFIOT: Data fusion for Internet of Things", J. Network and Systems Management, vol. 28, no. 4, pp. 1136-1160, 2020.

[7] S. Gite and H. Agrawal, "On context awareness for multisensor data fusion in IoT", Proc. Second Int. Conf. Computer and Communication Technologies: IC3T, vol. 3, Springer India, 2016.

[8] A. M. Sharafaddini, K. K. Esfahani, and N. Mansouri, "Deep learning approaches to detect breast cancer: A comprehensive review", Multimedia Tools and Applications, vol. 83, no. 5, pp. 1-112, 2024.

[9] Hurtado, J. V., & Valada, A. "Chapter 12 - Semantic Scene Segmentation for Robotics", Deep Learning for Robot Perception and Cognition, edited by A. Iosifidis and A. Tefas, Academic Press, 2022, pp. 279-311. ISBN 9780323857871. https://doi.org/10.1016/B978-0-32-385787-1.00017-8.

[10] J. F. Ferreira, et al., "Sensing and artificial perception for robots in precision forestry: A survey", Robotics, vol. 12, no. 5, pp. 139, 2023.

[11] A. Akbar, et al., "Real-time probabilistic data fusion for large-scale IoT applications", IEEE Access, vol. 6, pp. 10015-10027, 2018.

[12] W. Ding, et al., "A survey on data fusion in the Internet of Things: Towards secure and privacy-preserving fusion", Information Fusion, vol. 51, pp. 129-144, 2019.

[13] H. Lee, B. Lee, K. Park, and R. Elmasri, "Fusion techniques for reliable information: A survey", Int. J. Digital Content Technology and its Applications, vol. 4, pp. 74-88, 2010.

[14] F. Alam, R. Mehmood, I. Katib, N. N. Albogami, and A. Albeshri, "Data fusion and IoT for smart ubiquitous environments: A survey", IEEE Access, vol. 5, pp. 9533-9554, 2017.

[15] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art", Information Fusion, vol. 14, pp. 28-44, 2013.

[16] F. Kibrete, D. E. Woldemichael, and H. S. Gebremedhen, "Multi-sensor data fusion in intelligent fault diagnosis of rotating machines: A comprehensive review", Measurement, vol. 174, pp. 114658, 2024.

[17] D. Sadhukhan, S. Ray, and M. Dasgupta, "Data fusion in Internet of Medical Things: Towards trust management, security, and privacy", Data Fusion Techniques and Applications for Smart Healthcare, Academic Press, 2024, pp. 281-297.

[18] N. E. I. Hamda, A. Hadjali, and M. Lagha, "Multisensor data fusion in IoT environments in Dempster-Shafer theory setting: An improved evidence distance-based approach", Sensors, vol. 23, no. 11, pp. 5141, 2023.

[19] F. Mandreoli and M. Montangero, "Dealing with data heterogeneity in a data fusion perspective: Models, methodologies, and algorithms", Data Handling in Science and Technology, vol. 31, Elsevier, 2019, pp. 235-270.

[20] Gao, J., et al., "A survey on deep learning for multimodal data fusion", Neural Computation, vol. 32, no. 5, pp. 829-864, 2020.

[21] W. Cao, W. Feng, Q. Lin, G. Cao, and Z. He, "A review of hashing methods for multimodal retrieval", IEEE Access, vol. 8, pp. 15377-15391, 2020.

[22] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and

taxonomy", IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423-443, 2018.

[23] A. Biswas and H.-C. Wang, "Autonomous vehicles enabled by the integration of IoT, edge intelligence, 5G, and blockchain", Sensors, vol. 23, no. 4, pp. 1963, 2023.

[24] N. U. A. Tahir, et al., "Object detection in autonomous vehicles under adverse weather: A review of traditional and deep learning approaches", Algorithms, vol. 17, no. 3, pp. 103, 2024.

[25] M. Pawłowski, A. Wróblewska, and S. Sysko-Romańczuk, "Effective techniques for multimodal data fusion: A comparative analysis", Sensors, vol. 23, no. 5, pp. 2381, 2023.

[26] S. Mu, M. Cui, and X. Huang, "Multimodal data fusion in learning analytics: A systematic review", Sensors, vol. 20, no. 23, pp. 6856, 2020.

[27] I. Gallo, A. Calefati, and S. Nawaz, "Multimodal classification fusion in real-world scenarios", Proc. 14th IAPR Int. Conf. Document Analysis and Recognition (ICDAR), vol. 5, IEEE, 2017.

[28] G. Muhammad, et al., "A comprehensive survey on multimodal medical signals fusion for smart healthcare systems", Information Fusion, vol. 76, pp. 355-375, 2021.

[29] L. Jiang and C. Wu, "A massive multi-modal perception data classification method using deep learning based on Internet of Things", Int. J. Wireless Inf. Networks, vol. 27, no. 2, pp. 226-233, 2020.

[30] J. Gao, et al., "A survey on deep learning for multimodal data fusion", Neural Computation, vol. 32, no. 5, pp. 829-864, 2020.

[31] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives", IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798-1828, 2013.

[32] S. El-Sappagh, et al., "Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data", Neurocomputing, vol. 412, pp. 197-215, 2020.

[33] A. B. Said, et al., "Multimodal deep learning approach for joint EEG-EMG data compression and classification", Proc. 2017 IEEE Wireless Communications and Networking Conf. (WCNC), IEEE, 2017.

[34] P. T. Nguyen, et al., "Deep learning based optimal multimodal fusion framework for intrusion detection systems for healthcare data", Computers, Mater. & Continua, vol. 66, no. 3, pp. 2437-2455, 2021.

[35] A. Tsanousa, et al., "A review of multisensor data fusion solutions in smart manufacturing: Systems and trends", Sensors, vol. 22, no. 5, pp. 1734, 2022.

[36] A. John, et al., "Multimodal multiresolution data fusion using convolutional neural networks for IoT wearable sensing", IEEE Trans. Biomed. Circuits Syst., vol. 15, no. 6, pp. 1161-1173, 2021.

[37] Y. Zhang, et al., "Deep multimodal fusion for semantic image segmentation: A survey", Image Vision Comput., vol. 105, pp. 104042, 2021.

[38] W. Mucha and M. Kampel, "Depth and thermal images in face detection—a detailed comparison between image modalities", Proc. 2022 5th Int. Conf. Machine Vision and Applications (ICMVA), 2022.

[39] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481-2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

