Usage of Intermediate Fusion of Multimodal Data for Dangerous Objects Enhanced Detection*

Hlib Shchur[1,†] and Iryna Dumyn[1,*,†]

[1] Lviv Polytechnic National University, Bandery 12, 79000, Lviv, Ukraine

Abstract

In real-world applications like autonomous driving, maritime navigation, and industrial monitoring, reliably detecting dangerous objects is critical. Traditional object detection systems that rely on just one type of sensor often struggle when conditions are challenging — whether due to adverse weather, low light, or when objects are only partially visible. In this study, the last publications explore innovative multimodal sensor fusion techniques. These studies combine information from cameras, LiDAR, thermal, terahertz, and tactile sensors to create detection systems that are both more accurate and more robust. Building on these insights, the current paper aims to propose a unified framework that merges visual and sensory data using an intermediate-level fusion strategy enhanced by attention mechanisms. The proposed approach extracts detailed features from each sensor and fuses them into a single, cohesive representation. It also introduces an object criticality score - considering factors like distance, relative velocity, and orientation to prioritize high-risk objects. A hypothetical example shows how the system might work in practice.

Keywords

Multimodal Sensor Fusion, Object Detection, Autonomous Systems, LiDAR, Attention Mechanisms, Feature Extraction, Risk-Based Detection

1. Introduction

Ensuring the reliable detection of dangerous objects is absolutely critical in many real-world settings — from autonomous vehicles maneuvering through busy urban streets and maritime vessels navigating treacherous waters to industrial facilities and smart waste management systems in crowded cities. Traditional object detection systems[1], which typically rely on just one sensor type (like standard RGB cameras), often struggle under challenging conditions such as low-light environments, occlusions, or adverse weather. These shortcomings have sparked a growing interest in multimodal sensor fusion[2], where data from diverse sensors — such as cameras, LiDAR, thermal/infrared sensors, tactile sensors, and even terahertz imaging — are combined to deliver a more robust and reliable detection performance.

The process of demining areas contaminated with explosive devices remains one of the most pressing global challenges[3]. According to international organizations, a significant portion of land in conflict zones and post-war regions remains affected by landmines, posing a serious threat to civilian populations and hindering economic development. The application of robotic systems in this field represents a promising direction, as it significantly reduces risks for deminers, enhances demining efficiency, and shortens the duration of operations. Autonomous and remotely operated demining robots can function in challenging conditions, including high-threat areas and difficult-to-access terrains.

Despite these advantages, traditional control of demining robots in real-world environments faces several challenges. Key issues include limited situational awareness of the operator due to delays in video signal and data transmission, difficulties in navigating uneven terrain, and potential

*SMARTINDUSTRY-2025: 2nd International Conference on Smart Automation & Robotics for Future Industry, April 03 -05, 2025, Lviv, Ukraine

[1,*] Corresponding author.

[†] These authors contributed equally.

hlib.o.shchur@lpnu.ua (H. Shchur); iryna.b.shvorob@lpnu.ua (I. Dumyn)

0000-0002-3796-2866 (H. Shchur); 0000-0001-5569-2647 (I. Dumyn)

errors in decision-making algorithms that may lead to mission failures. Additionally, real-world testing of demining robots requires substantial financial resources and specially designed testing grounds, limiting their evaluation across a wide range of scenarios.

A promising approach to addressing these challenges is the use of virtual modeling for testing and training operators of robotic systems. Modern simulation platforms enable the creation of realistic environments where demining scenarios can be practiced, various threats can be modeled, and autonomous control algorithms can be adapted[4]. This approach significantly reduces testing costs, improves operator training, and facilitates more effective learning in safe conditions. Furthermore, the integration of machine learning methods in virtual environments enhances the adaptability of robotic systems to dynamic changes in real-world conditions. Multimodal data plays a key role in complex analysis and decision-making systems, as it combines information from different sensors (e.g., images, lidar data, audio, temperature readings), which increases the accuracy and reliability of processing. Relational and non-relational databases, graph structures, and specialized platforms for streaming large amounts of data are used to efficiently store such data[5]. Processing multimodal data includes deep learning, graph algorithms for establishing relationships between different modalities, and various data fusion methods.

This paper proposes a unified framework that fuses visual and sensory modalities to enhance the detection of dangerous objects. By leveraging intermediate-level fusion with attention mechanisms, proposed approach combines feature representations from multiple sensors and integrates an object criticality model to prioritize detections based on safety relevance. The framework is designed to be robust across a variety of environments and applicable to multiple domains, ultimately addressing the limitations of single-modality systems and advancing the state of safety-critical detection systems.

2. Literature Review

A careful examination of recent literature reveals a rich variety of approaches and challenges in the field of multimodal sensor fusion for object detection. The review of the latest scientific investigations in the area of object detection is provided in this section.

Thompson [6] delves into the realm of maritime object detection by combining LiDAR and vision data. In his study, high-fidelity GPS/INS information is fused with 3D LiDAR point clouds and camera images to track and classify objects on autonomous surface vehicles. The result is a detection system that achieves an impressive 98.7% accuracy across six object classes. However, the study also highlights important challenges — sensor alignment, accurate coordinate transformation, and the creation of reliable occupancy grids — which are crucial for extracting objects in the ever-changing maritime environment.

Vadidar et al. [7] focus on overcoming the limitations of conventional RGB cameras by fusing visual and thermal (infrared) data for autonomous driving. Their unified learning pipeline, centered around an innovative RGB-thermal (RGBT) fusion network, leverages an entropy-block attention module (EBAM) to refine the feature fusion process. This attention-based approach results in a notable 10% improvement in mean Average Precision (mAP) over existing methods, making it a powerful solution for reliable object detection under low-light or adverse weather conditions.

Bhown [8] tackles the critical issue of long-range detection for autonomous trucks, which must identify vulnerable road users (VRUs) in time to avoid collisions. Large vehicles require extended detection ranges because of their slower maneuverability compared to smaller cars. By fusing data from LiDAR and monocular cameras, Bhown's method compensates for the inherent sparsity of LiDAR point clouds at long distances. This fusion strategy is essential for ensuring that large autonomous vehicles can detect objects in urban and suburban environments where space is limited and reaction time is critical.

In the context of smart city applications, Alsubaei et al. [9] address the challenge of detecting and classifying small objects for effective garbage waste management. Their work leverages an enhanced version of the RefineDet deep learning model, with hyperparameters optimally tuned

using an arithmetic optimization algorithm (AOA). Although the focus is on waste segregation, the techniques developed have broader implications for detecting small, dangerous objects in complex environments, demonstrating the versatility of their approach.

Ceccarelli and Montecchi [10] provide a critical analysis of traditional object detection metrics, arguing that conventional measures like Average Precision do not adequately account for safety and reliability. They introduce an object criticality model that factors in an object's distance, relative velocity, and trajectory—elements that determine the potential risk posed by the object. This approach shifts the focus from merely detecting objects to prioritizing those that could significantly impact safety, a concept that is particularly relevant for autonomous driving systems.

Tabrik et al. [11] explore the intriguing overlap between visual and tactile perception. Their experiments with virtual 3D objects, or "digital embryos," reveal that both the visual and tactile systems share common shape features when it comes to object recognition. This finding suggests that the cognitive processes underlying these two sensory modalities are remarkably similar, which in turn supports the idea of integrating tactile data with visual data in robotic systems to enhance overall recognition performance.

Building on the interplay between vision and touch, Rouhafzay and Cretu [12] propose a framework in which visual attention guides tactile data acquisition. In their system, visually selected object contours determine where tactile data should be sequentially collected. By combining both cutaneous (surface) and kinesthetic (movement-based) cues through a deep learning approach employing CNNs, their framework achieves a very high recognition accuracy of 98.97%. This adaptive strategy mirrors how humans explore objects, and it demonstrates the benefits of a synergistic visuo-tactile approach.

Ahmad and Del Bue [13] present mmFUSION, an intermediate-level fusion framework that specifically addresses the challenges of integrating features from heterogeneous sensors like cameras and LiDAR. Their approach uses separate encoders to process each modality, and then employs cross-modality and multi-modality attention mechanisms to fuse these features effectively. The method not only preserves the detailed semantic and spatial information from each sensor but also achieves superior performance on standard benchmarks like KITTI and NuScenes.

Önal and Dandıl [14] take a slightly different approach by focusing on the detection of unsafe behaviors in workplace environments. Their system, Unsafe-Net, combines the spatial detection power of YOLO v4 with the temporal analysis capabilities of ConvLSTM networks. After processing 39 days of factory video footage, their hybrid approach achieves a classification accuracy of 95.81% and an action recognition latency of just 0.14 seconds. Although their primary application is workplace safety, the underlying techniques are highly relevant to object detection in other safety-critical domains.

Finally, Danso et al. [15] explore the use of terahertz imaging for detecting concealed dangerous objects, a method particularly useful for security screening applications. Terahertz images, despite being safe and non-ionizing, are often plagued by low resolution and noise. To address these issues, the authors enhance the YOLOv5 model with a BiFPN module and employ transfer learning to fine-tune the network. Their incremental improvements in mAP metrics highlight the potential of combining non-traditional imaging modalities with deep learning for detecting hidden hazards.

3. Proposed Approach

Below, the proposed unified fusion framework that integrates visual and sensory modalities to improve threat detection is described. The proposed approach is designed around a mid-level fusion strategy that uses attention mechanisms to dynamically weight and combine features from various sensors. This section details the overall architecture, key processing steps, and rationale for the proposed approach.

The framework is structured in multiple stages that transform raw sensor data into a consolidated detection decision.

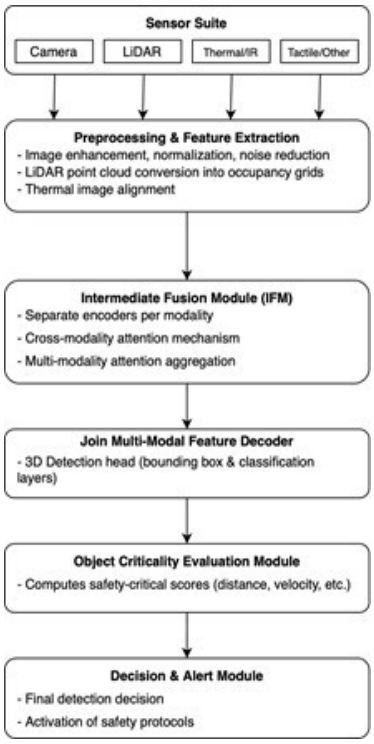

Figure 1: Complete data flow from sensor input to safety-critical output.

The sensor suite provides raw data from multiple modalities. Preprocessing extracts features which are then fused in the Intermediate Fusion Module using attention mechanisms. The fused features are decoded into object detections that are further evaluated for safety-criticality before triggering the final decision and alerting systems.

The detailed list of the main modules of the proposed framework is provided below.

Module 1 — Sensor Suite and Preprocessing. The framework starts with a Sensor Suite that includes:

1. **Camera** – This sensor captures RGB images, which are important for obtaining semantic details and texture information [7, 13].
2. **LiDAR** – This sensor provides accurate depth and spatial data that is later transformed into 3D occupancy grids [6, 8].
3. **Thermal/IR Sensor** – This sensor records temperature gradients, which helps in detecting objects in low-light conditions or during adverse weather [7].
4. **Tactile Sensors (Optional)** – These sensors collect cutaneous and kinesthetic data, which are useful for analyzing the shape and texture of objects [12, 11].

Each sensor's raw data is processed through specific preprocessing steps:

1. **Visual Data:** The data is enhanced and normalized using CNN-based methods to reduce noise and improve contrast.
2. **LiDAR Data:** The data is converted into structured formats, such as voxel grids or occupancy maps, to aid in feature extraction.
3. **Thermal Data:** The data is synchronized with camera frames to ensure spatial alignment between the different modalities.
4. **Tactile Data:** The data is transformed into feature maps that capture cues related to surface pressure and texture.

Module 2 — Intermediate Fusion Module. At the core of the proposed framework lies the Intermediate Fusion Module (IFM), which is responsible for combining features from different

modalities into a common representation. Unlike early fusion — which concatenates raw data — and late fusion — which aggregates final decisions, intermediate fusion leverages high-level features while preserving spatial and semantic integrity.

The IFM consists of two main steps, listed below.

Step 1. Separate Encoding. Each modality's features are encoded into a lower-dimensional space while maintaining key geometric and semantic details. This is accomplished using modality-specific encoders:

$$f_i = E_i(S_i), \tag{1}$$

where each f_i is the extracted feature from sensor i, S_i is the sensor input and E_i represents the encoder for modality i.[13].

Step 2. Attention-Based Fusion. To adaptively weight the contribution of each modality, is proposed to use a two-stage attention mechanism.

1. Cross-Modality Attention:

For each modality, a score S_i is computed that reflects its relevance in the current context:

$$a_i = \frac{e^{s_i}}{\sum_{j=1}^{N} e^{s_j}}, \tag{2}$$

This softmax operation normalizes the scores, ensuring that the attention weights a_i sum to one.

2. Aggregation:

The final fused feature F is a weighted sum of the individual modality features:

$$F = \sum_{i=1}^{N} a_i * f_i, \tag{3}$$

This process, inspired by the approach in mmFUSION [13], effectively integrates complementary information and mitigates the deficiencies of any single modality.

Module 3 — Joint Feature Decoding and Object Detection. Once the features have been fused, the Joint Multi-Modal Feature Decoder processes the combined feature vector F to produce object detection outputs. This stage typically involves a 3D detection head that predicts:

- **Bounding Boxes — l**ocalization of objects in 3D space;
- **Object Classes — c**lassification labels based on learned features;
- **Confidence Scores — p**robability estimates for each detection.

The detection process can be represented as:

$$Output = D(F), \tag{4}$$

where D is the decoder function that translates fused features into actionable detection information [8].

Module 4 — Object Criticality Evaluation. For safety-critical applications, not all detections are equally important. The Object Criticality Evaluation Module assigns a criticality score C to each detected object based on parameters such as:

- **Distance** (d) from the sensor.
- **Relative Velocity** (v) towards the platform.
- **Orientation** (θ) of the object relative to the collision path.

A representative formula for calculating the criticality score is:

$$C = e^{-\alpha d} * v * \theta, \tag{5}$$

where α is a decay constant that modulates the influence of distance [10]. This scoring mechanism ensures that objects posing a higher risk (e.g., closer and on a collision course) are prioritized in the final decision-making process.

Module 5 — Decision and Alert Module. The final stage combines the raw detection confidence $S_{detection}$ from the joint decoder with the criticality score C to determine whether an object is deemed dangerous. This is computed as:

$$S_{final} = \beta * S_{detection} + (1-\beta) * C, \quad (6)$$

where β ($0 \le \beta \le 1$) is a weighting factor balancing detection confidence and criticality [10]. If S_{final} exceeds a predefined threshold, the system issues a detection decision and triggers corresponding safety protocols.

The proposed framework integrates multiple sensor modalities at an intermediate level to exploit the strengths of each sensor while mitigating their individual weaknesses. The process begins with dedicated preprocessing and feature extraction from raw sensor data, followed by an attention-based fusion that produces a robust, unified feature representation. A joint decoder then translates these features into object detections, which are further evaluated for safety-criticality. Finally, a decision module synthesizes this information to yield a final detection outcome and, if necessary, initiate safety alerts.

By adopting this framework, systems can achieve enhanced detection accuracy and robustness in various complex and dynamic environments, thus making them more suitable for applications where safety is of utmost importance.

4. Hypothetical Example

In this section, the example of the operation of the proposed intermediate fusion framework through a detailed, hypothetical scenario is illustrated. The example demonstrates how multiple sensor inputs are processed, fused, and evaluated to make a safety-critical detection decision.

4.1. Scenario Description

Imagine an autonomous truck operating in an urban environment approaching an intersection. The system is tasked with detecting a pedestrian who is potentially crossing the road in an unsafe manner. The detection process involves three sensor modalities:

- Camera (RGB) captures visual information, including color and texture, to identify objects.
- LiDAR provides depth information by generating point clouds, crucial for estimating object distance.
- Thermal/IR Sensor captures temperature differences, which can highlight living beings even under low-light conditions.

4.2. Sensor Inputs and Preprocessing

For this example, assume the following sensor observations:

- Camera detects a candidate pedestrian with a raw confidence score of 0.90. After image enhancement and feature extraction (using a CNN encoder), the extracted feature is denoted as f_{cam}.
- LiDAR returns a sparse point cloud corresponding to an object with a raw confidence score of 0.80. The LiDAR data is converted into an occupancy grid and then processed by its dedicated encoder to produce feature f_{LiDAR}.
- Thermal/IR Sensor detects a warm signature in the same region with a confidence score of 0.85. Thermal features are extracted after alignment with the camera frame, resulting in feature $f_{thermal}$.

4.3. Intermediate Fusion Process

Each sensor's feature is weighted according to its relevance, as determined by the attention mechanism. Assume that under the current environmental conditions (e.g., dusk with low ambient light), the system assigns the following attention weights:

1. Camera: $a_{cam}=0.6$
2. LiDAR: $a_{LiDAR}=0.3$
3. Thermal: $a_{thermal}=0.1$

The fused feature F is computed as below:

$$F = a_{cam}*f_{cam}+a_{LiDAR}*f_{LiDAR}+a_{thermal}*f_{thermal}, \quad (7)$$

Simultaneously, the raw detection confidences from each modality are fused (as a simplified weighted sum) to produce an overall detection confidence:

$$S_{detection}=(0.6*0.90)+(0.3*0.85)+(0.1*0.85)=0.54+0.24+0.085=0.86 \quad (8)$$

4.4. Object Criticality Evaluation

Given the safety-critical context, the system computes an object criticality score C to prioritize objects based on potential risk.

For the current example, suppose the following values are measured or estimated:

1. Distance, d : 10 meters from the truck.
2. Relative Velocity, v : The pedestrian is moving toward the truck at 2 m/s.
3. Orientation Factor, θ : The pedestrian's path is directly toward the truck (θ = 1).
4. Decay Constant, α : 0.1, chosen to modulate the impact of distance.

The criticality score is then calculated as:

$$C=e^{-\alpha d}*v*\theta=e^{-0.1*10}*2*1=e^{-1}*2\approx 0.3679*2=0.7358\,, \quad (9)$$

4.5. Final Detection Decision

The final detection score S_{final} is derived by combining the fused detection confidence and the criticality score. Using a weighting factor $\beta=0.7$ (to prioritize raw detection confidence with still considering safety-critical information):

$$S_{final}=\beta*S_{detection}+(1-\beta)*C, \quad (10)$$

Substituting the values:

$$S_{final}=0.7*0.865+0.3*0.7358\approx 0.6055+0.2207=0.8262 \quad (11)$$

Assume the system's detection threshold is set at 0.80. Since $S_{final}=0.8262$ exceeds this threshold, the system classifies the object as dangerous. Consequently, safety protocols are activated - such as issuing and audible and visual alert to initiate braking or evasive maneuvers.

Below is an illustration of the proposed structure that integrates multiple sensor modalities at an intermediate level to leverage the strengths of each sensor while reducing their individual weaknesses. The process begins with specialized preprocessing and feature extraction from raw sensor data, followed by an attention-based fusion that produces a robust unified feature representation. Next, a joint decoder transforms these features into object detection, which are then evaluated for safety-criticality. Finally, a decision-making module synthesizes this information to generate the final detection result and, if necessary, trigger safety alerts.

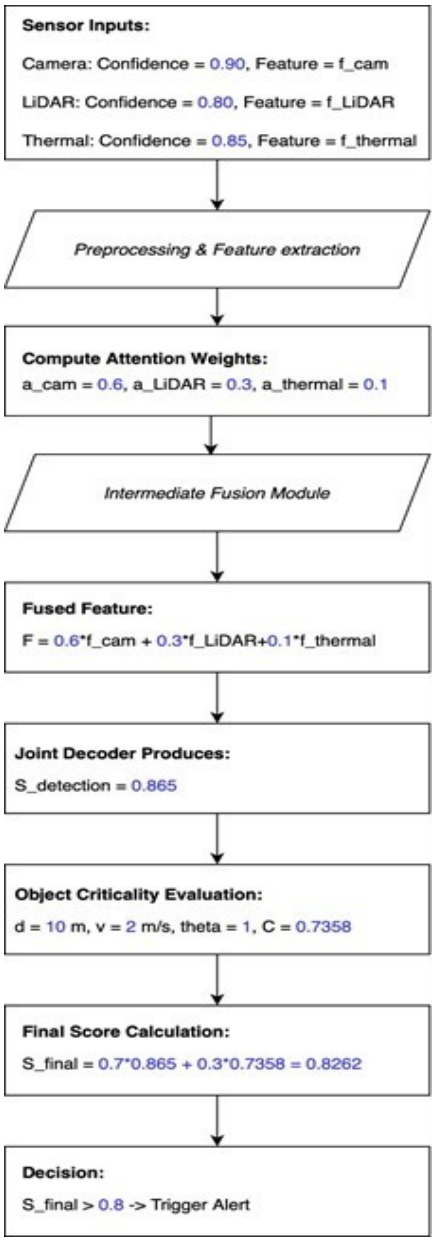

Figure 2: Hypothetical Example Data Flow.

This flow diagram outlines the sequential process from sensor input through to the final decision, demonstrating how the intermediate fusion and safety-critical evaluation work together.

4.6. Summary

This hypothetical example illustrates the comprehensive process of the proposed fusion framework. Initially, raw sensor data from various sources is preprocessed and encoded to prepare it for further analysis. Following this, an attention-based fusion mechanism is applied to dynamically weight and combine features, resulting in a unified representation. The framework then performs a safety-critical assessment by computing object criticality based on factors such as distance, relative velocity, and orientation. Finally, the detection confidence is combined with the criticality score to determine whether the object is hazardous.

The example demonstrates that, even when sensor confidence and conditions vary, the proposed framework can robustly integrate multimodal data to enhance the detection of safety-critical objects. This approach is particularly beneficial in real-world scenarios, where the timely identification of dangerous objects is essential.

5. Discussion

The proposed fusion framework and accompanying mathematical formulations address several long-standing challenges in detecting dangerous objects across varied, real-world scenarios. In this section, the benefits, limitations, and future prospects of the proposed approach are discussed.

5.1. Advantages

The framework improves robustness and accuracy by combining several sensor modalities such as RGB cameras, LiDAR, thermal sensors, and tactile data. This design takes advantage of the strengths of each sensor. For instance, RGB cameras provide detailed semantic information, while LiDAR offers precise depth measurements. Thermal sensors work effectively in low-light conditions, and tactile data adds useful insights into object shape and texture. This blend of sensor inputs enhances overall detection accuracy and reliability, especially in challenging situations where systems using a single modality might fail.

The system also uses an attention-based fusion mechanism that adjusts the weight of each sensor based on the current environment. For example, in poor lighting or bad weather, the system can give more importance to thermal or LiDAR data than to RGB images. The softmax-based attention mechanism helps to ensure that the most reliable sensor inputs have the greatest influence on the final feature representation.

Additionally, the framework includes an object criticality model to increase safety by prioritizing detections that present higher risks. By combining detection confidence with factors such as distance, speed, and orientation, the system focuses quickly on objects that may be on a collision path. This approach is vital in areas like autonomous driving and maritime navigation, where detection errors can have serious consequences.

Finally, the framework uses an intermediate fusion strategy that avoids the problems of both early fusion, which can lead to misaligned raw data, and late fusion, which may depend on inaccurate initial proposals. By merging high-level features from each sensor, the approach maintains important semantic and spatial details, leading to better detection performance.

5.2. Limitations

Ensuring precise calibration between sensors is one of the most challenging aspects of multimodal fusion. Differences in field of view, resolution, and sampling rates may lead to misalignment that reduces the quality of integrated features. Although the framework applies preprocessing and synchronization steps, further research is needed to develop more robust and adaptive calibration methods.

Another issue is computational complexity. The use of attention mechanisms and intermediate fusion increases the computational overhead, which can be a significant challenge for real-time applications such as autonomous vehicles and industrial monitoring systems. Optimizing the network architecture and utilizing hardware accelerators like GPUs or TPUs may help mitigate these costs.

Furthermore, many of the current datasets used for object detection in safety-critical domains are limited in diversity and lack comprehensive multimodal labeling. This scarcity of fully annotated multimodal datasets makes it difficult to completely train and evaluate advanced fusion models. Expanding these datasets to cover a wider range of dangerous objects and adverse conditions is essential for future progress.

5.3. Future Work

Future research should concentrate on developing dynamic calibration techniques that automatically align sensor data in real time. These techniques may use adaptive algorithms that adjust to changes in sensor positioning and environmental conditions. Addressing the computational overhead is also critical for real-time applications; exploring model compression,

efficient network architectures, and specialized hardware solutions can help bridge the gap between theoretical research and practical deployment. Moreover, there is an urgent need for comprehensive datasets containing synchronized and labeled data from multiple sensor modalities across various scenarios. Such datasets would enable more thorough training, benchmarking, and refinement of multimodal fusion models, ultimately improving their applicability in real-world settings. Finally, although the current framework focuses on detection, future systems might integrate these outputs with higher-level decision-making and control processes. For example, in autonomous vehicles, detection results could be directly linked to trajectory planning algorithms that make immediate adjustments to prevent collisions.

5.4. Summary

The discussion underscores that the proposed fusion framework effectively addresses key challenges in dangerous object detection by leveraging multimodal sensor data and advanced fusion strategies. The dynamic weighting through attention mechanisms and the inclusion of a safety-critical evaluation component significantly enhance detection robustness and reliability. Nonetheless, challenges remain in sensor calibration, computational efficiency, and dataset availability. Addressing these limitations through future research will be essential to fully realize the potential of multimodal sensor fusion in safety-critical applications.

References from earlier sections consistently emphasize the importance of robust data fusion, dynamic sensor weighting, and safety-critical performance evaluation. The current work builds upon these foundational ideas, providing a comprehensive, adaptable, and practical solution for enhanced detection in complex environments.

6. Conclusion

In this paper, a comprehensive framework for the enhanced detection of dangerous objects through the fusion of visual and sensory modalities was presented. By synthesizing insights from the latest studies, the proposed approach addresses key challenges encountered in safety-critical applications such as autonomous driving, maritime navigation, and industrial monitoring.

The described unified framework employs an intermediate-level fusion strategy that leverages dedicated encoders for each sensor modality — such as cameras, LiDAR, thermal sensors, and tactile sensors — to extract high-level features. These features are dynamically weighted using attention mechanisms and fused into a unified representation, which is then decoded to produce robust 3D object detections. A critical component of the proposed approach is the object criticality model, which quantifies the risk posed by detected objects based on their distance, relative velocity, and orientation. This enables the system to prioritize high-risk objects, thus enhancing safety in environments where timely detection is essential.

The hypothetical example in Section 4 further illustrates how the proposed framework effectively integrates multimodal sensor data to produce reliable detection decisions in a real-world scenario. While the proposed framework shows significant promise, challenges remain. Accurate sensor calibration, computational efficiency for real-time processing, and the need for expanded, well-annotated multimodal datasets are areas that warrant further investigation. Future work should focus on developing dynamic calibration methods, optimizing the fusion architecture, and integrating the detection module with higher-level decision-making systems to enable seamless real-time responses.

In summary, the integration of visual and sensory modalities through intermediate fusion and attention mechanisms represents a powerful solution for detecting dangerous objects in complex, dynamic environments. Current approach starts the future research and practical implementations in safety-critical domains, ultimately contributing to the development of next-generation autonomous systems with enhanced robustness and reliability.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] C. Arya, A. Tripathi, P. Singh, M. Diwakar, K. Sharma, and H. Pandey, Object detection using deep learning: a review, Journal of Physics: Conference Series, 1854(1), p. 012012, 2021. doi:10.1088/1742-6596/1854/1/012012.

[2] F. Zhao, C. Zhang, and B. Geng, Deep Multimodal Data Fusion, ACM Comput. Surv., 56(9), Article 216, pp. 1-36, 2024. doi:10.1145/3649447.

[3] H. Fedorenko, H. Fesenko, and V. Kharchenko, Analysis of methods and development of a concept for guaranteed detection and recognition of explosive objects, Innovative Technologies and Scientific Solutions for Industries, 4(22), pp. 20-31, 2022.

[4] H. Shchur, Intelligent system for demining robot control in a virtual environment, Herald of Khmelnytskyi National University. Technical Sciences, 335(3(1)), pp. 326-329, 2024. doi:10.31891/2307-5732-2024-335-3-43.

[5] I. Dumyn, O. Basystiuk, and A. Dumyn, Graph-based approaches for multimodal medical data processing, 2023.

[6] D. J. Thompson, Maritime Object Detection, Tracking, and Classification Using LiDAR and Vision-Based Sensor Fusion, 2017.

[7] M. Vadidar, A. Kariminezhad, C. Mayr, L. Kloeker, and L. Eckstein, Robust Environment Perception for Automated Driving: A Unified Learning Pipeline for Visual-Infrared Object Detection. In 2022 IEEE Intelligent Vehicles Symposium (IV) (pp. 367-374). IEEE.

[8] A. Bhown, Improving Long-Range 3D Object Detection Methods for Autonomous Box Trucks Using Sensor Fusion. 2022.

[9] F. S. Alsubaei, F. N. Al-Wesabi, and A. M. Hilal, Deep Learning-Based Small Object Detection and Classification Model for Garbage Waste Management in Smart Cities and IoT Environment, 2022.

[10] A. Ceccarelli and L. Montecchi, Evaluating Object (mis)Detection from a Safety and Reliability Perspective: Discussion and Measures. IEEE Access, 11, 44952-44963. 2023.

[11] S. Tabrik, M. Behroozi, L. Schlaffke, S. Heba, M. Lenz, S. Lissek, O. Güntürkün, H. R. Dinse, and M. Tegenthoff, Visual and Tactile Sensory Systems Share Common Features in Object Recognition. Eneuro, 8(5). 2021.

[12] G. Rouhafzay and A.-M. Cretu, An Application of Deep Learning to Tactile Data for Object Recognition under Visual Guidance. Sensors, 19(7), 1534. 2019.

[13] J. Ahmad and A. Del Bue, mmFUSION: Multimodal Fusion for 3D Objects Detection.arXiv preprint arXiv:2311.04058. 2023.

[14] O. Önal and E. Dandıl, Unsafe-Net: YOLO v4 and ConvLSTM Based Computer Vision System for Real-Time Detection of Unsafe Behaviours in Workplace. Multimedia Tools and Applications, 1-27, 2024.

[15] S. A. Danso, L. Shang, D. Hu, J. Odoom, Q. Liu, and B. Nyarko, Hidden Dangerous Object Recognition in Terahertz Images Using Deep Learning Methods. Applied Sciences, 12(15), 7354, 2022.