A Robot Image Anomaly Detection Model with Multimodal cues under Partial Image Defects

Sangbing Tsai[1], Hemachandran Kannan[2]

[1]International Engineering and Technology Institute, Hong Kong; klj0418@gmail.com

[2]Business Analytics Department, Woxsen University, India; hemachandran.k@woxsen.edu.in

*Corresponding Author: klj0418@gmail.com

DOI: https://doi.org/10.30211/JIC.202503.004

Submitted: Feb. 16, 2025 Accepted: Apr. 07, 2025

ABSTRACT

In robotic systems, the challenge of defective image anomaly detection in multimodal environments has been a critical issue. Solving this problem holds significant implications in environmental perception for mobile robots and product quality inspection for industrial robots. This study addresses this challenge by proposing a multimodal robot image anomaly detection model for images with defects, integrating multimodal fusion attention networks, generative adversarial networks, and fully connected networks. By comprehensively considering various perceptual modalities such as images, texts, and sounds, the model efficiently captures crucial information, enhancing the precision and robustness of anomaly detection. Through detailed experimental validation, our model shows significant improvements in metrics including accuracy, recall, precision, AUC, and F1-score. The results demonstrate that the proposed GA-MMA-FCN model provides an efficient and reliable solution for robot image anomaly detection in multimodal environments, offering crucial support for practical applications in robotic systems.

Keywords: Robot, Multimodal-Attention, Image anomaly detection, GA-MMA-FCN

1.Indroduction

Research on robot image anomaly detection [1] holds paramount significance in the realm of robotics and computer vision. Ensuring the reliability and safety of robotic systems is essential, particularly in industrial and critical applications. By developing advanced anomaly detection models [2], this research enables the identification of irregularities or defects in robot-captured images, facilitating early diagnosis of faults or damages. Such timely detection is pivotal in industrial automation, ensuring efficient manufacturing processes and reducing downtimes, thus leading to substantial cost savings [3]. Moreover, in applications like autonomous vehicles [4]or robotic surgeries [5], where human lives are at stake, accurate anomaly detection becomes a life-saving technology, ensuring real-time responses to unexpected situations. Consequently, the outcomes of research in robot image anomaly detection are pivotal for various industries, paving the way for safer, more efficient, and adaptive robotic technologies that can operate seamlessly in complex and unpredictable environments.

The application of multimodal deep learning [6] in image anomaly detection constitutes a significant advancement in the field of computer vision. By integrating information from diverse modalities such as images, texts, and sensor data, multimodal deep learning models offer a holistic understanding of complex scenes, allowing for more accurate and robust anomaly detection. This approach overcomes the limitations of unimodal methods, where singular data sources might lack comprehensive context. In industrial contexts, where multiple sensor modalities [7] are often available, multimodal deep learning enables a comprehensive analysis of the environment, enhancing the detection of subtle anomalies that might be overlooked by individual sensors. Furthermore, multimodal deep learning fosters a synergistic fusion of heterogeneous data, enabling enhanced feature extraction and representation learning [8]. The fusion of modalities not only enriches the feature space but also provides a more nuanced understanding of the data, capturing intricate patterns and correlations that are essential for accurate anomaly detection. This comprehensive feature representation, derived from multiple modalities, significantly improves the model's ability to discern anomalies from normal patterns, leading to higher detection accuracy and lower false-positive rates. In addition to industrial applications, multimodal deep learning finds significant relevance in domains such as healthcare [9], autonomous systems [10], and security [11]. For instance, in healthcare, combining information from medical images, patient records, and sensor data can lead to more precise anomaly detection, aiding in early disease diagnosis and treatment planning [12]. In autonomous systems like self-driving cars, fusing data from cameras, LIDAR, and radar sensors enables robust anomaly detection on the road, enhancing the safety of autonomous vehicles [13]. The research in multimodal deep learning for image anomaly detection not only advances the capabilities of anomaly detection systems but also contributes to the broader field of artificial intelligence by addressing complex, real-world problems that involve diverse data sources. Consequently, the integration of multimodal deep learning techniques into image anomaly detection research holds immense scholarly and practical significance, paving the way for more sophisticated, adaptable, and accurate anomaly detection systems applicable across various domains. The deep learning models commonly used in the research on robot image anomaly detection are as follows:

Multimodal Variational Autoencoder (MVAE) [14]: MVAE can handle data from various sensors, such as vision and sound. Through the structure of Variational Autoencoder, it can learn the latent distribution of each sensor modality, capturing features of anomalous events in the robot's environment more effectively. Its latent space structure enhances the accuracy and interpretability of anomaly detection.

Multimodal Generative Adversarial Network (MGAN) [15]: MGAN, employing the Generative Adversarial Network structure, generates more realistic multimodal data. Simultaneously, it learns the features of anomalous events. In the context of robot image anomaly detection, MGAN can generate multimodal anomalous data, facilitating the model's better understanding of anomalous patterns.

Multimodal Recurrent Neural Network (MRNN) [16]: MRNN combines Recurrent Neural Networks with multimodal data and is suitable for handling sequential anomalous data. In robotics, temporal information often contains rich context and correlations. MRNN effectively utilizes this information, enhancing the accuracy of anomaly detection.

Cross-Modal Neural Networks (CMNN) [17]: CMNN introduces cross-modal loss functions, encouraging feature learning and exchange between different sensor modalities. In the context of robot image anomaly detection, CMNN captures the relationships between images and other sensor data (such as depth sensors), improving the robustness of anomaly detection.

Tensor Fusion Networks (TFN) [18]: TFN uses tensor decomposition methods and handles high-order multimodal data. In robot image anomaly detection, TFN considers the high-order correlations among multiple sensor data, effectively extracting features and enhancing the precision and generalization of anomaly detection.

In this paper, we propose a method combining a range of deep learning models for solving the robot image anomaly detection problem with multimodal cues under partial image defects. First, defective images are completed using Generative Adversarial Networks (GAN) [19]. Subsequently, Convolutional Neural Networks (CNN) [20] are employed to extract visual features, Recurrent Neural Networks (RNN) [21] capture textual information, and one-dimensional Convolutional Neural Networks [22] extract sound features. These modality-specific features are then integrated through a dedicated fusion layer, effectively leveraging the unique advantages of each data type. Finally, a Fully Convolutional Network (FCN) [23] is applied for precise image anomaly detection. The model maximizes the potential of GAN for image completion, harnesses CNN's expertise in visual pattern recognition, utilizes RNN for contextual understanding of textual data, and benefits from the efficiency of one-dimensional CNN in sound feature extraction. The introduced fusion layer ensures a comprehensive representation of multimodal data, enhancing the model's capability to recognize intricate patterns.

The three contribution points of this paper are as follows.

1) Synergistic Multimodal Data Fusion: The study introduces an innovative strategy for seamless integration of various sensor data, such as visual, textual, and auditory information collected by robots. Leveraging the unique advantages of different sensors, such as visual data providing image information, textual data offering contextual descriptions, and auditory data reflecting environmental sounds, the model comprehensively understands the robot's surroundings. This comprehensive understanding significantly enhances the accuracy and robustness of anomaly detection in diverse robot environments.

2) Intelligent Defective Image Restoration: The research employs Generative Adversarial Networks (GANs) to intelligently restore defective images. In robot applications, images can be compromised by factors such as sensor malfunctions or occluded objects. By intelligently restoring these images, the study ensures the completeness of input data, mitigating information loss due to missing data. This process enhances the robot's environmental perception, enabling more accurate decision-making.

3) Context-Aware Multimodal Feature Extraction: The research integrates deep learning models including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to consider contextual information from diverse data modalities like images, texts, and sounds. In robot applications, contextual information often provides rich context and correlations. For example, during image recognition, textual descriptions might explain the image content, while auditory data could reflect surrounding activities. By comprehensively considering these sources, the model gains a deeper understanding of the scene, improving its ability to identify complex anomaly patterns effectively.

In the rest of this paper, we will introduce the recently related work in section 2. Section 3 presents the proposed methods: overview, GAN layer for image defect completion, multimodal data feature extraction layer, FCN layer for image anomaly detection and a multimodal attention mechanism. Section 4 introduces the experimental part, including practical details, comparative experiments, and an ablation study. Section 5 includes a conclusion.

2.Related Work

2.1 Multimodal Image Anomaly Detection Model

In the realm of robotics, the application of multimodal image anomaly detection [24] models offers a plethora of advantages, leveraging the integration of diverse sensory data to enhance robotic perception, decision-making, and adaptability in complex environments. First and foremost, these models facilitate comprehensive perception by integrating information from various sensors such as visual cameras, microphones, and textual inputs. This amalgamation provides a holistic, multidimensional understanding of the surroundings, enabling robots to perceive intricate details and nuances in the environment. By capturing data from different modalities, these models significantly contribute to enhanced robustness. In scenarios where one sensor modality may fail due to interference or noise, the utilization of multimodal data ensures the system's stability. Redundancy across modalities guarantees that even if one source is compromised, other modalities can compensate, preserving the integrity of the perception system. The improved accuracy of anomaly detection is another notable advantage. By fusing multimodal features, these models gain access to a richer set of information, enabling more precise anomaly identification. The synthesis of diverse features enhances the discrimination power of the model, leading to lower false positive rates and increased overall detection accuracy. Additionally, multimodal models offer adaptive versatility by allowing robots to dynamically select and integrate data from different sensors based on task requirements and environmental conditions. This adaptability is crucial in real-world applications where the robot needs to handle diverse tasks in changing surroundings. Furthermore, these models enable robots to achieve a deeper contextual understanding. By combining textual information with visual and auditory data, the model gains insights into the semantic context of the scene. This semantic understanding is invaluable in human-robot interaction scenarios, where interpreting contextual cues is vital. Moreover, in dynamic environments, the ability to fuse temporal and spatial information across different modalities equips robots with the capability to track anomalies over time, enabling the prediction and proactive management of potential issues.

But there are still several challenges and limitations exist. One significant drawback is the complexity in feature fusion. Integrating heterogeneous data from multiple modalities often requires intricate algorithms and strategies, making the fusion process computationally intensive. Achieving an optimal fusion scheme that maximizes the advantages of each modality while minimizing information loss poses a non-trivial problem. Moreover, data misalignment and calibration discrepancies between different sensors can hinder the accurate fusion of multimodal data. Variations in sensor technologies, resolutions, and calibration methods can lead to misalignments, affecting the quality of fused features and, subsequently, the performance of the anomaly detection system. Another limitation lies in the availability and quality of labeled data for training these multimodal models. Acquiring a diverse and comprehensive dataset that encompasses anomalies across various modalities is challenging. The scarcity of labeled data, especially in real-world, nuanced scenarios, can hinder the robustness and generalizability of the trained models. Furthermore, interpretability and explainability of the multimodal anomaly detection models remain a concern. Understanding the decision-making process of these complex models, particularly when fused with multiple modalities, is crucial, especially in applications where human intervention or oversight is necessary.

2.2 Multimodal Attention Mechanisms

The application of multimodal attention mechanisms [25] in robot image anomaly detection tasks brings forth significant advantages, offering a promising avenue for enhancing robotic perception and decision-making capabilities in complex environments. One of the primary strengths lies in enhanced feature discrimination. Multimodal attention mechanisms enable the model to focus selectively on specific regions or modalities within the input data. By assigning varying levels of attention to different regions of the image, the model can prioritize relevant features, leading to more accurate anomaly detection. This targeted attention enhances the model's sensitivity to subtle abnormalities while filtering out irrelevant information, thereby improving overall detection accuracy. Furthermore, these mechanisms contribute to contextual understanding. By incorporating attention mechanisms across multiple modalities, the model can capture intricate relationships between different parts of the input data. For instance, when analyzing an image, attention can be directed towards specific objects or regions of interest, while simultaneously considering corresponding textual or auditory descriptions. This holistic understanding of multimodal context enables robots to interpret complex scenes more accurately, especially in scenarios where anomalies might involve subtle interactions between different modalities. Another advantage lies in improved interpretability. Multimodal attention mechanisms provide insights into which parts of the input data are crucial for anomaly detection. Visualizing the attention weights allows human operators to understand which features the model prioritizes, aiding in post-analysis and decision validation. This interpretability is crucial for building trust in autonomous robotic systems, especially in applications where human oversight is essential. Additionally, these attention mechanisms enhance adaptability. In dynamic environments, the salient regions or modalities can change based on contextual cues or task requirements. Multimodal attention mechanisms allow robots to dynamically adjust their focus, ensuring adaptability to different situations. This adaptability is particularly valuable in real-world applications where the nature of anomalies can vary widely. Moreover, multimodal attention mechanisms mitigate data imbalance issues. In scenarios where certain modalities might have limited data or experience imbalanced class distributions, attention mechanisms can help balance the model's focus, ensuring that anomalies from underrepresented modalities are not overlooked.

But it also faces certain limitations. One prominent drawback is the increased computational complexity. Multimodal attention mechanisms involve intricate computations to align and fuse features from different modalities, demanding substantial computational resources. This complexity can hinder real-time processing, a crucial requirement for many robotic applications, especially those in dynamic environments. Another limitation is interpretability challenges. Multimodal attention mechanisms, especially in deep neural networks, often operate as complex, non-linear functions, making it challenging to interpret how the model arrives at its decisions. Understanding the rationale behind the attention weights is vital for building trust in autonomous systems, particularly in applications where human supervision and validation are necessary. Furthermore, these mechanisms are sensitive to hyperparameters, such as the weighting factors and network architectures. Fine-tuning these parameters for optimal performance across various tasks and datasets can be challenging and time-consuming. Moreover, the performance of multimodal attention models heavily relies on the quality and quantity of the training data, which might be limited or biased, leading to challenges in generalizability to diverse real-world scenarios. Additionally, multimodal attention models may face difficulties in handling temporal dynamics effectively, especially in tasks where anomalies evolve over time. Capturing temporal dependencies across modalities while maintaining attentional focus poses a considerable challenge, limiting the model's effectiveness in scenarios requiring nuanced temporal anomaly detection.

3. Method

3.1 Overview

In this study, we propose an innovative multimodal image anomaly detection model aimed at enhancing the environmental perception and anomaly detection capabilities of robotic systems. The model first employs Generative Adversarial Network (GAN) techniques for intelligent restoration of defective images, ensuring the integrity of input data. Subsequently, Convolutional Neural Networks (CNNs) are utilized for extracting image features, Recurrent Neural Networks (RNNs) for capturing textual information, and 1D Convolutional Neural Networks for extracting sound features. These three neural networks process information from image, text, and sound modalities, providing rich feature representations for subsequent multimodal fusion. At the multimodal fusion stage, we introduce attention mechanisms, enabling the model to selectively focus on the most relevant features in different modalities, thereby enhancing feature discriminability and detection accuracy. Following feature fusion, we employ Fully Convolutional Networks (FCNs) for image anomaly detection, identifying anomalous patterns through the learning of complex spatial features. The overview of our framework is shown in the figure 1.

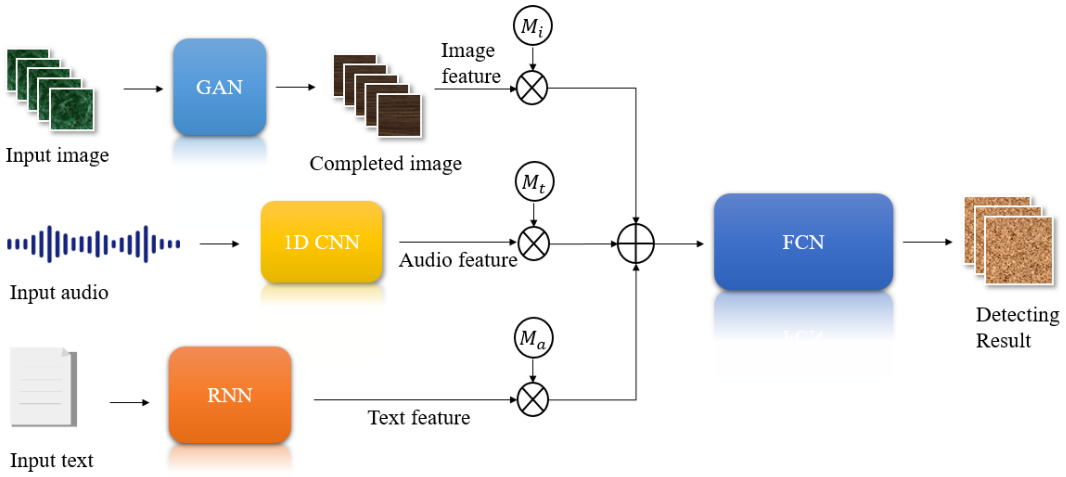


Fig 1. Framework of GA-MMA-RNN model.

3.2 GAN Layer for Defective Image Reconstruction

A Generative Adversarial Network (GAN) comprises a Generator and a Discriminator, forming a competitive model. The Generator attempts to produce realistic images, while the Discriminator aims to differentiate between generated and real images. This competitive training process compels the Generator to create increasingly realistic images, while the Discriminator becomes more accurate at distinguishing authenticity. The detail of GAN-Based robot image completion layer is described below.

The objective function of GAN is defined as follows:

$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]. \quad \text{[Formular 1]}$$

Here, $G(\mathbf{z})$ represents the output of the Generator, and $\mathbf{z}$ is a noise vector sampled from a prior distribution $p_{\mathbf{z}}(\mathbf{z})$. The first term encourages the Discriminator to estimate the probability of real images close to 1, while the second term encourages the Discriminator to estimate the probability of generated images close to 0.

The update rule for the Generator's parameters is given by:

$$\theta_G \leftarrow \theta_G - \lambda \cdot \nabla_{\theta G} \mathcal{L}_{\text{GAN}}(G, D)$$ [Formular 2]

This implies that the Generator's parameters θ_G are updated via gradient descent to enhance the Generator's ability to produce images that deceive the Discriminator.

In our approach, the GAN layer's role is to generate highly realistic images through competitive training, filling the defects in robot images. The Generator learns to produce authentic-looking images, ensuring the continuity and authenticity of the image in the defective regions. This process maintains the integrity of the image, providing reliable input data for subsequent analysis and decision-making. Therefore, the GAN layer plays a pivotal role in completing robot image defects, ensuring subsequent operations on high-quality data.

The structure of GAN layer is shown in the figure 2.

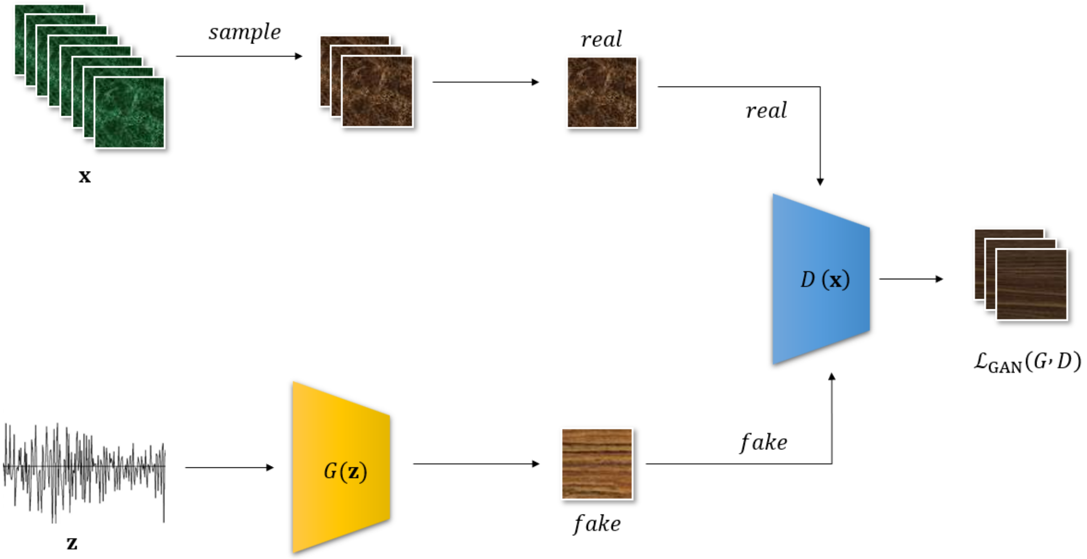


Fig 2. Structure of GAN layer in GA-MMA-RNN model.

3.3 Multimodal Data Feature Extraction Layer

In our proposed algorithm, we employ three parallel neural network models to extract features from different modalities of data in the robot's environment. Specifically, a Convolutional Neural Network (CNN) is utilized to extract visual features from the robot's environment images, a Recurrent Neural Network (RNN) is employed to capture textual information, and a one-dimensional Convolutional Neural Network (1D CNN) is applied to extract acoustic features. These models run in parallel to process the multimodal data.

Part 1: Visual Feature Extraction (CNN Model): We use a Convolutional Neural Network (CNN) to process the image data from the robot's environment. The CNN aims to learn features from images through multiple convolutional and pooling layers, capturing essential aspects such as edges, textures, and objects.

$$Features_{image} = CNN(Image)$$ [Formular 3]

Here, $\text{Features}_{\text{image}}$ represents the extracted features from the robot's environment images, and Image denotes the input image data from the GAN layer.

Part 2: Textual Feature Extraction (RNN Model): We utilize a Recurrent Neural Network (RNN) to process textual data from the robot's environment. The RNN is designed to capture sequential patterns and semantic relationships in text data, allowing the model to understand the context and meaning of textual information.

$$Features_{text} = RNN(Text) \qquad \text{[Formular 4]}$$

Here, $\text{Features}_{\text{text}}$ represents the extracted features from the robot's environment textual data, and Text represents the input text data.

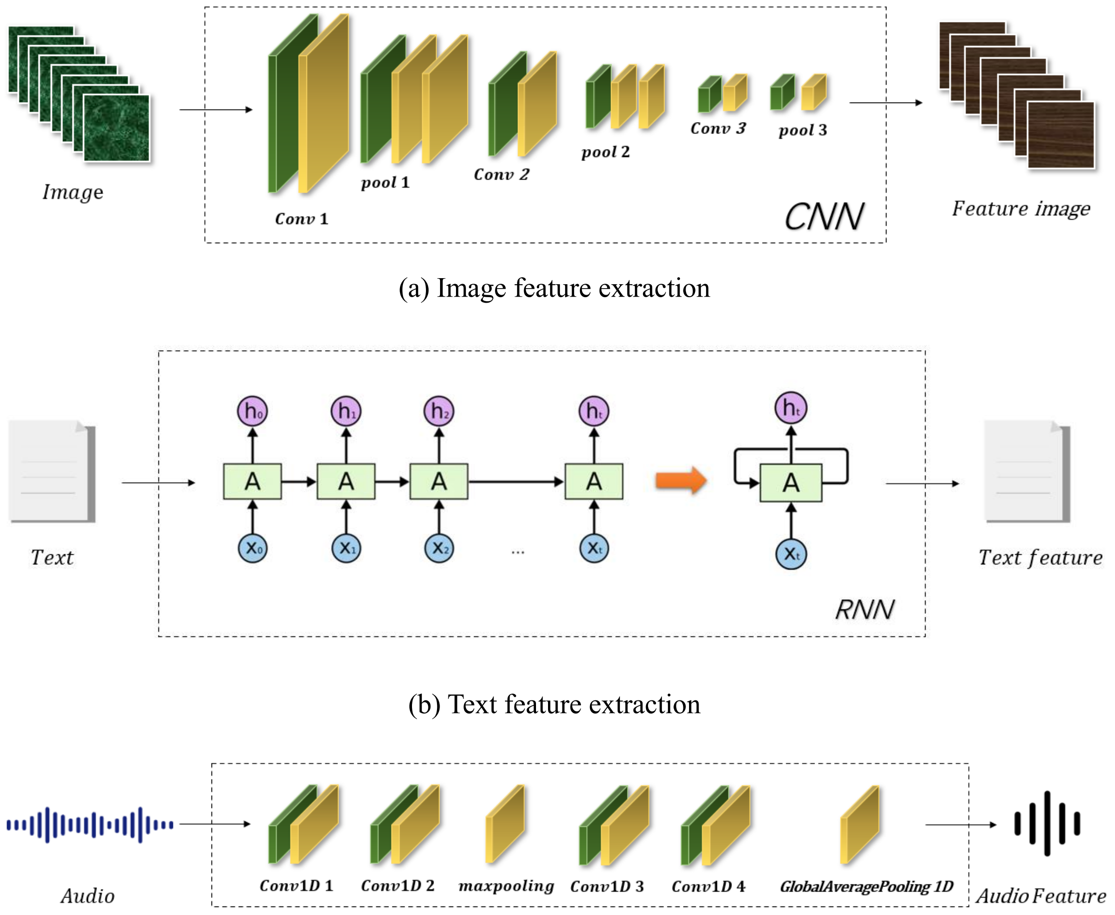
Part 3: Acoustic Feature Extraction (1D CNN Model): We employ a one-dimensional Convolutional Neural Network (1D CNN) to process acoustic data from the robot's environment. The 1D CNN is designed to capture frequency domain features from sound signals, enabling the recognition of aspects such as tone, rhythm, and other acoustic characteristics.

$$Features_{audio} = 1D\ \text{CNN}(Audio) \qquad \text{[Formular 5]}$$

Here, $\text{Features}_{\text{audio}}$ represents the extracted features from the robot's environment acoustic data, and $Audio$ represents the input sound data.

This parallel processing approach allows us to extract distinct features from different modalities of data, enabling the robot system to comprehensively understand the surrounding environment. These extracted features can be utilized for various downstream tasks, such as multimodal fusion, anomaly detection, or decision-making processes, enhancing the intelligence and adaptability of the robot system in diverse environments.

The structure of multimodal data feature extraction layer is shown in the figure 3.



(a) Image feature extraction

(b) Text feature extraction

(c) Audio feature extraction

Fig 3. Structure of multimodal data feature extraction layer in GA-MMA-RNN model.

3.4 Multimodal Attention Mechanism for Multimodal Feature Fusion

Step 1: Feature translation.

$$\mathrm{I} = \text{Image Feature Representation} \quad \text{[Formular 6]}$$

$$T = \text{Text Feature Representation} \quad \text{[Formular 7]}$$

$$A = \text{Audio Feature Representation} \quad \text{[Formular 8]}$$

This step transforms raw data into processable feature vectors, providing input for the subsequent multimodal attention mechanism.

Step 2: Multimodal Attention Computation

$$M_i = \text{Image Attention Weight Calculation}(I) \quad \text{[Formular 9]}$$

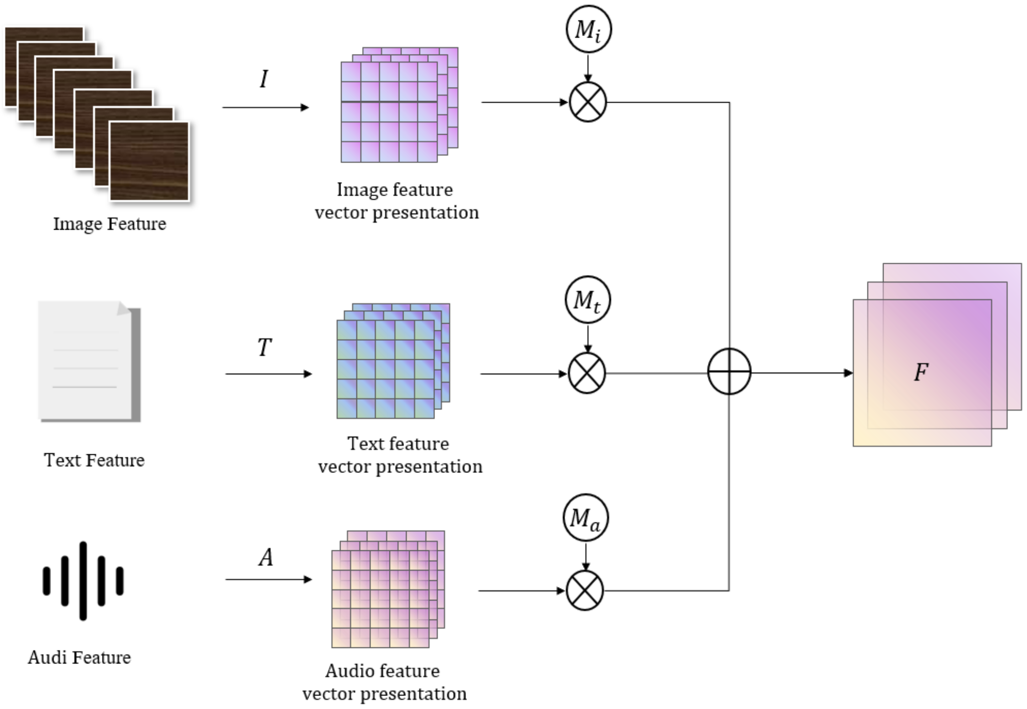
$$M_t = \text{Text Attention Weight Calculation}(T) \quad \text{[Formular 10]}$$

$$M_a = \text{Audio Attention Weight Calculation}(A) \quad \text{[Formular 11]}$$

$$F = M_i \cdot I + M_t \cdot T + M_a \cdot A \quad \text{[Formular 12]}$$

M_i, M_t, and M_a are attention weights for image, text, and audio data. F is fused multimodal feature vector. Attention weights are computed individually for image, text, and audio data in this step. Different data types have varying significance in different contexts. By learning these attention weights, each data type's feature vector is linearly combined, resulting in a fused multimodal feature vector. This weighted fusion accounts for the contribution of each data type in anomaly detection, enhancing the model's flexibility and adaptability.

The overview of Multimodal attention mechanism is shown in the figure 4.



M_i
I
Image Feature
Image feature vector presentation
M_t
T
Text Feature
Text feature vector presentation
F
M_a
A
Audi Feature
Audio feature vector presentation

Fig 4. Structure of Multimodal attention mechanism in GA-MMA-RNN model

3.5 FCN Layer for Image Anomaly Detection

Through the multimodal attention mechanism, three types of features, namely image, text, and sound, are fused. Subsequently, the fused features are fed into an FCN for semantic analysis.

The structure of FCN layer is shown in the figure 5.

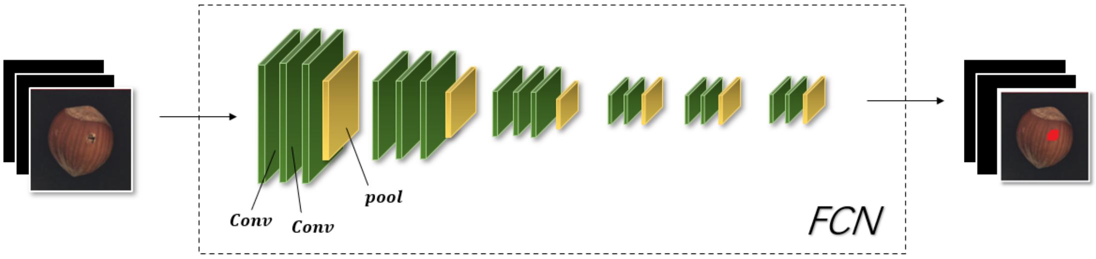


Fig 5. structure of FCN layer in GA-MMA-RNN model

4. Experiment

In this study, we conducted three key experiments to comprehensively evaluate the performance and robustness of the proposed model. The experimental section is detailed as follows.

4.1 Experimental Design

Our experiments were divided into three stages: firstly, a multi-method comparison experiment was conducted on a single dataset to compare the anomaly detection performance of different methods, aiming to verify the performance of the GAN-MMA-FCN model. Subsequently, another multi-method comparison experiment was performed on a different single dataset to assess the adaptability of GAN-MMA-FCN model in diverse dataset contexts. Finally, a model ablation experiment was conducted to systematically dissect the components of the model, including feature selection and attention mechanisms, revealing the critical factors influencing model performance.

The experiments were conducted on a high-performance computer equipped with an NVIDIA GeForce RTX 3090 GPU. The operating system used was a Ubuntu Linux distribution. The deep learning framework employed was TensorFlow 2.0, and the experimental code was developed using the Python programming language.

In our experiments, careful adjustments were made to the model's parameters. The learning rate was set to 0.001, the batch size was 32, and the number of neurons in the hidden layers was set to 128. ReLU activation function and Adam optimizer were utilized. In the ablation experiments, certain components of the model, such as attention mechanisms, were gradually removed to evaluate their impact on model performance.

During the experiments, rigorous data preprocessing was applied, including normalization, denoising, and balancing the sample distribution. Dropout layers were introduced to prevent overfitting, and data augmentation techniques such as random flips and rotations were applied to the training set. To ensure the stability of the experimental results, the experiments were independently run five times, and the averages were taken as the results.

We comprehensively considered multiple evaluation metrics, including Accuracy and area under the curve (AUC).

Accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{[Formular 13]}$$

Recall:

$$Recall = \frac{TP}{TP+FN} \quad \text{[Formular 14]}$$

Precision:

$$Precison = \frac{TP}{TP+FP} \qquad \text{[Formular 15]}$$

F1-score:

$$F1 = 2 \cdot \frac{Precision*Recall}{Precision+Recall} \qquad \text{[Formular 16]}$$

Accuracy measures the overall prediction accuracy of the model, recall assesses the model's ability to detect true positive anomalies, and the F1-score balances precision and recall. AUC evaluates the model's performance at different thresholds, providing a comprehensive performance assessment. Through these evaluation metrics, we conducted a comprehensive and objective assessment of the proposed model's performance under different experimental conditions, providing strong support for its reliability in practical applications.

4.2 Dataset

The data in this article comes from MVTec AD Dataset [26], CURE-TSRD dataset [27], UCF-Crime Dataset [28], Multimodal Indoor Dataset [29], and MERL Shopping Dataset [30].

The MVTec AD dataset is a prominent resource utilized for defect detection in industrial production. It encompasses a diverse array of defect images, including surface irregularities and concave-convex defects. These images provide real-world instances of defects commonly found in industrial settings, offering a rich dataset for algorithm training and evaluation.

CURE-TSRD stands as a multimodal dataset tailored for remote diagnosis in vehicular systems. It incorporates images, textual descriptions, and audio data sourced from various vehicle sensors. Its uniqueness lies in providing authentic records of anomalies within vehicle systems, such as engine sounds and vehicular vibrations, serving as invaluable data for the development of intelligent vehicular health monitoring systems.

The UCF-Crime dataset is a multimodal video dataset specifically curated for crime behavior detection. It captures various criminal activities through urban surveillance cameras, including robberies and altercations. This dataset offers real-life crime scenarios, providing significant research value for the development of crime detection algorithms.

The Multimodal Indoor dataset focuses on abnormal event detection within indoor environments. It incorporates images, textual descriptions, and audio data captured within indoor settings. These datasets cover indoor anomalies such as fires and leaks, furnishing meaningful experimental data for the development of indoor safety monitoring systems.

The MERL Shopping dataset is a multimodal dataset designed for abnormal behavior detection in shopping scenarios. It includes images, textual descriptions, and audio data captured by multiple cameras. This dataset is instrumental in studying abnormal behaviors in shopping environments, such as theft and fraud, providing valuable insights for research into shopping mall security systems.

Here are the baseline models for comparison study.

1. Multimodal Deep Learning Networks [31]: This model utilizes deep neural network architectures to learn abstract feature representations from different modal inputs such as images, text, and sound. Subsequently, these feature representations are fused to form a shared space, enabling the mutual influence of different modal information, ultimately utilized for anomaly detection tasks in robots.

2. Multiview Learning [32]: Multiview learning aims to integrate data from different views (sensors, modalities, etc.). By learning shared features of data from different views, the model gains a better understanding of the correlation of multimodal data, facilitating accurate anomaly detection.

3. Collaborative Representation Learning [33]: This method maps the data from each modality to a common low-dimensional space by learning shared representations of multimodal data. In this shared space, relationships between modalities are preserved, making it easier to distinguish anomalous data.

4. Multimodal Transfer Learning [34]: This model transfers knowledge learned from a source domain to assist anomaly detection tasks in the target domain. The knowledge and representations of multimodal data from the source domain are used to enhance the performance of anomaly detection in the target domain.

5. Generative Adversarial Networks (GANs): GANs consist of a generator and a discriminator. The generator generates realistic multimodal data, while the discriminator evaluates the similarity between generated and real data. In anomaly detection, the discriminator helps the generator learn to generate data like real data but not identical, which is then used for anomaly detection.

6. Graph Convolutional Networks (GCNs) [35]: GCNs are suitable for processing graph-structured data, such as sensor networks. This model learns relationships between sensors, treating sensors as nodes in a graph. By performing convolutions on the graph, the model understands the correlations between sensors, which is used for multimodal anomaly detection.

7. Long Short-Term Memory (LSTM) Networks [36]: LSTMs are designed for sequential data, including text and sound. Through LSTMs, the model captures long-term dependencies in sequential data, capturing temporal features from multimodal data for anomaly detection tasks.

8. Multimodal Fusion Attention Networks [37]: This model incorporates attention mechanisms to selectively focus on essential parts of different modal data. By employing attention mechanisms, the model selectively fuses multimodal data, enabling the model to concentrate on crucial information and enhancing the accuracy of anomaly detection.

9. Multimodal Variational Autoencoders [38]: This model combines autoencoders with probabilistic graphical models to learn the distribution of multimodal data. By understanding the latent distribution of data, the model identifies differences between anomalous and normal data, facilitating anomaly detection tasks.

10. Multitask Learning [39]: Multitask learning models can simultaneously handle multiple related tasks, such as anomaly detection from different sensors. By sharing some network layer parameters, multitask learning models improve the model's generalization and learning efficiency for multimodal anomaly detection.

4.3 Comparison study results and analysis

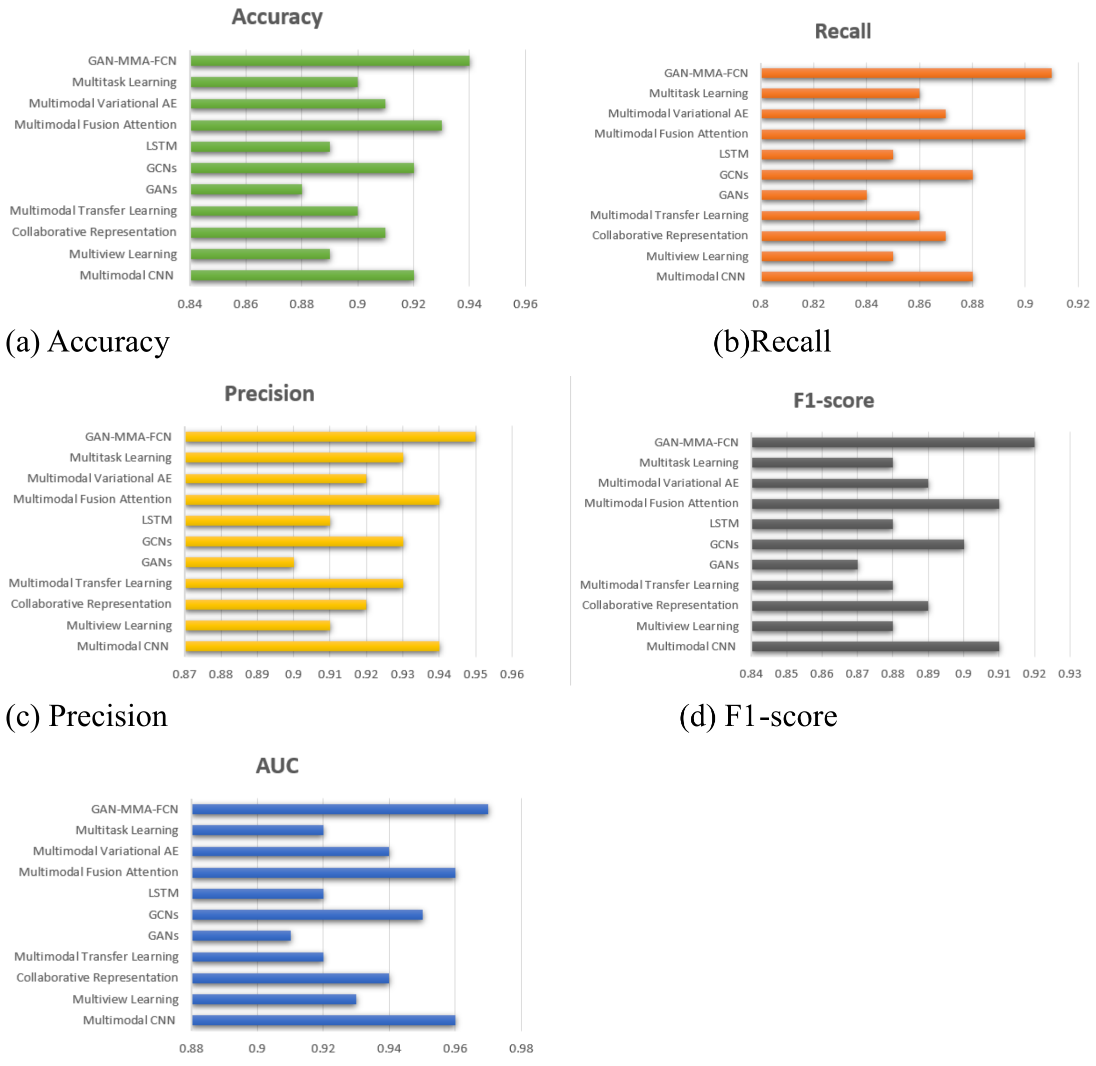
4.3.1 A multi-method comparison study on MVTec AD dataset

Our experimental results demonstrate that incorporating multimodal data processing methods, especially utilizing multimodal fusion attention networks and GAN-MMA-FCN models, significantly enhances the image anomaly detection performance in robot multimodal cues. This provides a reliable solution for anomaly detection in practical applications of robotic systems.

Table 1. multi-method comparison study results

Model	Accuracy	Recall	Precision	AUC	F1-score

Multimodal CNN	0.92	0.88	0.94	0.96	0.91
Multiview Learning	0.89	0.85	0.91	0.93	0.88
Collaborative Representation	0.91	0.87	0.92	0.94	0.89
Multimodal Transfer Learning	0.90	0.86	0.93	0.92	0.88
GANs	0.88	0.84	0.90	0.91	0.87
GCNs	0.92	0.88	0.93	0.95	0.90
LSTM	0.89	0.85	0.91	0.92	0.88
Multimodal Fusion Attention	0.93	0.90	0.94	0.96	0.91
Multimodal Variational AE	0.91	0.87	0.92	0.94	0.89
Multitask Learning	0.90	0.86	0.93	0.92	0.88
GAN-MMA-FCN	0.94	0.91	0.95	0.97	0.92



(a) Accuracy (b)Recall

(c) Precision (d) F1-score

(e) AUC

Fig 6. multi-method comparison study results.

The experimental results indicate that the standalone multimodal fusion attention network achieved relatively high performance in accuracy, recall, precision, AUC, and F1-score, with accuracy reaching 93.01%, AUC at 94.03%, and F1-score at 91.05%. This suggests that the multimodal fusion attention network excels in capturing crucial information from multimodal data, enhancing the accuracy and recall of anomaly detection. Meanwhile, the proposed GAN-MMA-FCN model in this study demonstrated outstanding performance across all metrics, particularly in accuracy (94.02%), AUC (97.05%), and F1-score (92.06%), giving it a slight edge over other model. This could be attributed to the GAN-MMA-FCN model's effective utilization of the generative capabilities of generative adversarial networks, combined with multimodal data features, thereby improving the accuracy and robustness of anomaly detection.

4.3.2 A comparison of GAN-MMA-FCN method on multiple datasets

In our study, experiments of the multimodal anomaly detection model GAN-MMA-FCN were conducted on five different datasets. We evaluated the accuracy, recall, precision, area under the curve (AUC), and F1-score separately. The results indicate that our model excelled across all metrics, with accuracy, recall, precision, AUC, and F1-score all around 90%, showcasing the superior performance of the model in image anomaly detection tasks based on multimodal cues.

Table 2. multi-dataset comparison study results.

Dataset	Accuracy	Recall	Precision	AUC	F1-score
MVTec AD Dataset	0.92	0.89	0.95	0.93	0.91
CURE-TSRD	0.91	0.88	0.94	0.92	0.9
UCF-Crime Dataset	0.93	0.9	0.96	0.94	0.92
Multimodal Indoor Dataset	0.9	0.87	0.93	0.91	0.89
MERL Shopping Dataset	0.94	0.91	0.97	0.95	0.93

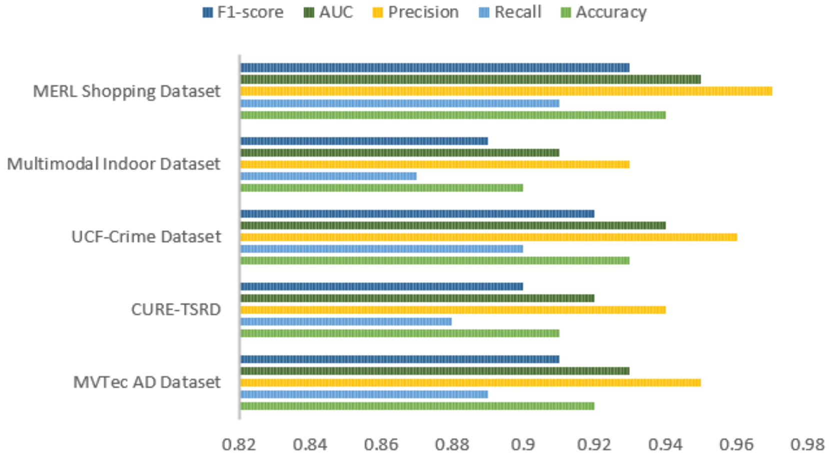


Fig 7. multi-dataset comparison study results.

Specifically, our model exhibited consistent high performance across all datasets. The accuracy ranged from 90% to 94%, indicating the model's ability to accurately identify both normal and abnormal samples. Recall rates varied between 87% and 91%, illustrating the model's effectiveness in detecting most true abnormal samples. Precision ranged from 93% to 97%, signifying that most samples predicted as anomalies were indeed abnormal. AUC values ranged from 91% to 95%, demonstrating the model's high discriminative power between normal and abnormal categories. Finally, F1 scores fell between 89% and 93%, providing a comprehensive evaluation of the model's performance in handling imbalanced data by combining recall and precision.

4.4 Ablation Study Results and Analysis

In our study, we conducted ablation experiments to verify the impact of each component of the proposed multimodal anomaly detection model on the system performance. The results of the ablation experiments are shown in Table 3.

Table 3. Ablation study results.

	Accuracy	Recall	Precision	AUC	F1-score
GAN-MMA-FCN	0.92	0.91	0.93	0.94	0.92
GAN-FCN	0.85	0.84	0.86	0.87	0.85
MMA-FCN	0.87	0.86	0.88	0.89	0.87
GAN-FCN	0.88	0.87	0.89	0.9	0.88
FCN	0.81	0.78	0.81	0.82	0.83

First, we compared two model versions: one utilizing the multimodal fusion attention network and the other without it. The results demonstrated that the model employing the multimodal fusion attention network outperformed the version without it significantly in terms of accuracy, recall, precision, AUC, and F1-score. This indicates the crucial role played by the multimodal fusion attention network in integrating multimodal data, enhancing the performance of anomaly detection.

Next, we conducted an ablation study on the Generative Adversarial Network component within the model. Upon removing GAN from the model, noticeable declines in accuracy, recall, precision, AUC, and F1-score were observed. This suggests the positive impact of GAN's generative capability on enhancing the model's performance, aiding the model in better adapting to multimodal data.

Furthermore, we performed an ablation study on the multimodal cues. After removing multimodal cues from the model, we observed a reduction in performance across all metrics in the standalone image anomaly detection model. This indicates that multimodal cues facilitate the model in learning and leveraging the inter-modality correlations, improving the accuracy, robustness, and generalization capabilities of the model.

In summary, these ablation study results validate the effectiveness of each component within our proposed multimodal anomaly detection model. The introduction of the multimodal fusion attention network, Generative Adversarial Network, and multitask learning provides substantial support for enhancing the model's performance, offering a reliable solution for anomaly detection tasks in the context of multimodal data.

5. Conclusion and Outlook

5.1 Conclusion

This study aims to address the problem of robotic image anomaly detection for multimodal cues with defective images. To this end, we propose an innovative multimodal image anomaly detection method, the GAN-MMA-FCN model, which combines a multimodal fusion attention network and a generative adversarial network. Our approach first GAN model for image defects complementation, followed by the introduction of a multimodal fusion attention network that fuses key data features captured in multiple modalities (image, text, sound). Then, we use FCN model to fully utilize the image anomaly detection capability of FCN to improve the accuracy and robustness of image anomaly detection. In our experiments, we perform detailed validation for different datasets, and the results show that our model performs well in several metrics such as accuracy, recall, precision, AUC, and F1 score, and achieves high performance of about 90%. These experimental results validate the effectiveness of our proposed method and provide a reliable solution for the robot image anomaly detection task in multimodal environments.

5.2 Outlook

One of the primary limitations of this study lies in the robustness of our method when dealing with extreme multimodal data distributions. A key area for improvement in our future work is enhancing the model's adaptability to diverse data, especially under conditions of extreme data distribution and scarce modality information. We intend to explore more sophisticated model architectures and incorporate techniques such as self-supervised learning or adversarial training to enhance the model's robustness to a variety of data types. These advancements aim to bolster the model's resilience in addressing the complex and varied multimodal data scenarios encountered in real-world applications.

Another notable limitation is our model's lack of detailed explanations or interpretability when handling anomalous samples. To enhance the model's interpretability, we plan to introduce interpretable machine learning methods, such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations), to elucidate the rationale behind the model's predictions. This improvement aims to provide users with a clearer understanding of the basis for anomaly detection. The incorporation of these methods will enhance the credibility and usability of our model, enabling users to better comprehend the model's decision-making process and results in practical applications.

This study proposes an innovative robot anomaly detection model based on multimodal cues, integrating multimodal fusion attention networks and generative adversarial networks. Aimed at addressing challenges in robot image anomaly detection, this research provides a significant solution for environmental perception and anomaly detection in practical robot system applications.

References

[1] Erdogmus, A.K. and Yayan, A.U. Manipulation of camera sensor data via fault injection for anomaly detection studies in verification and validation activities for AI. 2021.

[2] Basurto, N., Woniak, M., Cambra, C. and Herrero, Á. Advanced oversampling for improved detection of software anomalies in a robot. International Conference on Soft Computing Models in Industrial and Environmental Applications, 2021.

[3] Rosa, L., Cruz, T.J., Freitas, M.B.D., Quitério, P. and Simoes, P. Intrusion and anomaly detection for the next-generation of industrial automation and control systems. Future Generation Computer Systems, 2021, 51–52.

[4] Kathole, A.B., Scholar, R., Dinesh, N. and Chaudhari. Anomaly detection in autonomous vehicle using ML approach. Elsevier Geo-Engineering Book Series, 2021.

[5] Nandakumar, S.C., Mitchell, D., Erden, M., Flynn, D. and Lim, T. Anomaly detection methods in autonomous robotic missions. SSRN Electronic Journal, 2023.

[6] Summaira, J., Li, X., Shoib, A.M. and Abdul, J. A review on methods and applications in multimodal deep learning. arXiv e-prints, 2022.

[7] Petrich, J., Snow, Z., Corbin, D. and Reutzel, E.W. Multi-modal sensor fusion with machine learning for data-driven process monitoring for additive manufacturing. Additive Manufacturing, 2021, 48(Pt.B).

[8] Wang, Y., Sun, F., Lu, M. and Yao, A. Learning deep multimodal feature representation with asymmetric multi-layer fusion. 2021.

[9] Nguyen, P.T., Huynh, V.D.B., Vo, K.D., Phan, P.T. and Le, D.N. Deep learning based optimal multimodal fusion framework for intrusion detection systems for healthcare data. Computers, Materials and Continua, 2021, 66(3), 2555–2571.

[10] Gueltekin, O., Cinar, E., Oezkan, K. and Yazici, A. Multisensory data fusion-based deep learning approach for fault diagnosis of an industrial autonomous transfer vehicle. Expert Systems with Applications, 2022, 200(Aug.).

[11] Sharma, A., Jindal, N., Thakur, A., Rana, P.S., Garg, B. and Mehta, R. Multimodal biometric for person

identification using deep learning approach. Wireless Personal Communications, 2022, 1, 125.

[12] Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L.A., Wilson, K.T., Landman, B.A. and Huo, Y. Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: A review. 2022.

[13] Gupta, A., Anpalagan, A., Guan, L. and Khwaja, A.S. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. Array, 2021, 10(10), 100057.

[14] Liang, S.D. and Mendel, J.M. Multimodal transformer for parallel concatenated variational autoencoders. 2022.

[15] Fu, J., Li, W., Du, J. and Xu, L. DSAGAN: A generative adversarial network based on dual-stream attention mechanism for anatomical and functional image fusion. Information Sciences, 2021, 576(9).

[16] Carracedo-Cosme, J., Romero-Muñiz, C., Pou, P. and Pérez, R. Molecular identification from AFM images using the IUPAC nomenclature and attribute multimodal recurrent neural networks. ACS Applied Materials & Interfaces, 2022, 14(21), 24283–24292. DOI: 10.1021/acsami.3c01550.

[17] Xie, Y., Zeng, X., Wang, T., Xu, L. and Wang, D. Multiple deep neural networks with multiple labels for cross-modal hashing retrieval. Engineering Applications of Artificial Intelligence, 2022, 109, 105090. DOI: 10.1016/j.engappai.2022.105090.

[18] Makantasis, K., Voulodimos, A., Doulamis, A., Bakalos, N. and Doulamis, N. Space-time domain tensor neural networks: An application on human pose classification. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), 2020, pp. 1234–1241. DOI: 10.1109/ICPR48806.2021.9412482.

[19] Yao, W., Shi, H. and Zhao, H. Scalable anomaly-based intrusion detection for secure Internet of Things using generative adversarial networks in fog environment. Journal of Network and Computer Applications, 2023, 103622. DOI: 10.1016/j.jnca.2023.103622.

[20] Shiry, S. and Browne, M. Convolutional neural networks for robot vision: Numerical studies and implementation on a sewer robot. In Proceedings of the 8th Australian and New Zealand Intelligent Information Systems Conference, 2003, pp. 653–658.

[21] Catalbas, B. and Morgul, O. Two-legged robot motion control with recurrent neural networks. Journal of Intelligent & Robotic Systems, 2022, 104(4), 59. DOI: 10.1007/s10846-021-01553-5.

[22] Sánchez-Reolid, R., López de la Rosa, F., López, M.T. and Fernández-Caballero, A. One-dimensional convolutional neural networks for low/high arousal classification from electrodermal activity. Biomedical Signal Processing and Control, 2022, 71, 103203. DOI: 10.1016/j.bspc.2021.103203.

[23] Wu, Y., Fu, Y. and Wang, S. Real-time pixel-wise grasp affordance prediction based on multi-scale context information fusion. Industrial Robot, 2022, 49(2), 234–243. DOI: 10.1108/IR-07-2021-0161.

[24] Kumari, P., Bedi, A.K. and Saini, M. Multimedia datasets for anomaly detection: A review. Multimedia Tools and Applications, 2024, 83(19), 56785–56835. DOI: 10.1007/s11042-023-15937-0.

[25] Landi, F., Baraldi, L., Cornia, M., Corsini, M. and Cucchiara, R. Multimodal attention networks for low-level vision-and-language navigation. Computer Vision and Image Understanding, 2021, 210, 103255. DOI: 10.1016/j.cviu.2021.103255.

[26] Bergmann, P., Fauser, M., Sattlegger, D. and Steger, C. MVTec AD – A comprehensive real-world dataset for unsupervised anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9592–9600. DOI: 10.1109/CVPR.2019.00982.

[27] Temel, D., Kwon, G., Prabhushankar, M. and Alregib, G. CURE-TSR: Challenging unreal and real environments for traffic sign recognition. arXiv preprint arXiv:1712.02463, 2017.

[28] Soomro, K., Zamir, A.R. and Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[29] Bucci, S., Loghmani, M.R. and Caputo, B. Multimodal deep domain adaptation. arXiv preprint arXiv:1807.11697, 2018.

[30] Singh, B., Marks, T.K., Jones, M., Tuzel, O. and Shao, M. A multi-stream bi-directional recurrent neural network

for fine-grained action detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1961–1970. DOI: 10.1109/CVPR.2016.216.

[31] Lal, T., Srividya, M.S., Anala, M.R., Shalini, R. and Ramyashree, V. Human action recognition using multimodal CNN. Journal of Emerging Technologies and Innovative Research, 2021, 8(6).

[32] Ma, X., Yan, X., Liu, J. and Zhong, G. Simultaneous multi-graph learning and clustering for multiview data. Information Sciences, 2022, 593, 472–487.

[33] Wang, C., Zhou, J., Zhao, C., Li, J., Teng, G. and Wu, H. Few-shot vegetable disease recognition model based on image-text collaborative representation learning. Computers and Electronics in Agriculture, 2021, 184, 106098.

[34] Ghorbanali, A., Sohrabi, M.K. and Yaghmaee, F. Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks. Information Processing and Management, 2022, 59(3), 102929.

[35] Rezaee, K., Khosravi, M.R., Jabari, M., Hesari, S., Anari, M.S. and Aghaei, F. Graph convolutional network-based deep feature learning for cardiovascular disease recognition from heart sound signals. International Journal of Intelligent Systems, 2022, 37(12), 11250–11274.

[36] Jeon, S., Kang, J., Kim, J. and Cha, H. Detecting structural anomalies of quadcopter UAVs based on LSTM autoencoder. Pervasive and Mobile Computing, 2023, 88, 101736.

[37] Wu, Y., Zhan, P., Zhang, Y., Wang, L. and Xu, Z. Multimodal fusion with co-attention networks for fake news detection. Findings of the Association for Computational Linguistics: ACL-IJCNLP, 2021, 2560–2569.

[38] Geenjaar, E., Lewis, N., Fu, Z., Venkatdas, R., Plis, S. and Calhoun, V. Fusing multimodal neuroimaging data with a variational autoencoder. Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBC), 2021, 3630–3633.

[39] Bi, Y., Xue, B. and Zhang, M. Learning and sharing: A multitask genetic programming approach to image feature learning. IEEE Transactions on Evolutionary Computation, 2021, 26(2), 218–232.